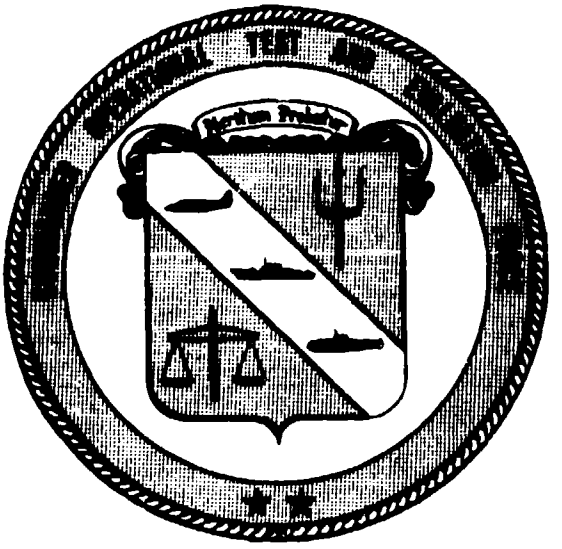
COMOPTEVFORINST 3960.7A

ANALYST NOTEBOOK



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A124 469	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Analyst Notebook		5. TYPE OF REPORT & PERIOD COVERED Handbook
		6. PERFORMING ORG. REPORT NUMBER COMOPTEVFORINST 3960.7A
7. AUTHOR(s) Commander Operational Test and Evaluation Force		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of the Navy Commander Operational Test and Evaluation Force Norfolk, VA 23511		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE 16 February 1982
		13. NUMBER OF PAGES 100
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Unlimited Distribution

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Test and Evaluation, Operational Test and Evaluation, Data Analysis, Error Analysis, Simulation, Reliability, Software Error Analysis, Statistical Analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This document is designed primarily for Project Analysts. As a supplement to standard analysis texts, it contains various examples and notes on data analysis techniques useful in operational test and evaluation.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DEPARTMENT OF THE NAVY
COMMANDER OPERATIONAL TEST AND EVALUATION FORCE
NORFOLK, VIRGINIA 23511

COMOPTEVFORINST 3960.7A
02B:ypc
16 February 1982

COMOPTEVFOR INSTRUCTION 3960.7A

Subj: Analyst's Notebook

Ref: (a) COMOPTEVFOR Instruction 3960.8, Project Analysis Guide

1. Purpose

a. This document provides guidance for various facets of OT&E (operational test and evaluation). It is designed primarily for Project Analysts. Operational Test Directors and Coordinators may find certain chapters pertinent.

b. As a supplement to standard analysis texts, this contains various examples and notes on analysis techniques useful in operational evaluations. The object is to share worthwhile techniques. Some examples are simple and straightforward; these serve as a refresher to a newly-arrived analyst. Other examples cover little known techniques or techniques that were recently presented in the literature that are expected to be useful at OPTEVFOR. The Headquarters Senior Analyst (Code 02B) has the original references on these latter techniques.

c. The examples and notes, listed in no particular sequence, pertain mainly to data analysis rather than to test planning. Test planning and experimental design in operational situations are covered to a great extent in reference (a), which is must reading for newly-arrived analysts.

d. Analysis in OT&E includes both art and science. Contributions by each analyst are necessary to share techniques.

2. Cancellation. This document cancels and supersedes COMOPTEVFORINST 3960.7.

H. A. FRENCH
Deputy Chief of Staff

Distribution:
COMOPTEVFORINST 5216.2C
List II, 1, 4

COMOPTEVFORINST 3960.7A
16 February 1982

Copy to:
DEPCOMOPTEVFORPAC (50)
CO, AIRTEVRON ONE (50)
CO, AIRTEVRON FOUR (50)
CO, AIRTEVRON FIVE (50)
OIC, OPTEVFORDET SUNNYVALE CA (10)
OIC, OPTEVFORDET PAX RIVER MD (4)
OIC, AIRTEVRON FIVE DET WHIDBEY ISLAND WA (6)
COMNAVOCEANSYSCEN (2)

Contents

Section 1

Monitoring Data Analysis

101. General. An important part of monitoring a project is keeping in touch if the data processing and analysis is being done by others. On scene involvement may save a month's time or effort. From an analysis point of view, here are some considerations to guide you during your visit.

a. Pre-Analysis

(1) Are computer processes being checked?
(2) Are data postings being checked?
(3) Are card inputs being verified?
(4) Are computer listings being examined?
(5) Are data being edited?
(6) Are specific rules being followed in editing?
(7) List of invalid runs prepared?
(8) List of out-of-line data prepared?
(9) Are all runs and all data being accounted for?
(10) Is data enclosure being prepared for report?
(11) Is table or scope of runs being prepared?
(12) Specific values of parameters used rather than zones?
(13) Sequence of runs available?
(14) Environmental conditions noted?
(15) Are data given tender loving care?
(16) Is processing time being minimized?

b. Analysis

(1) Does the Analysis Plan in the Test Plan apply?
(2) Should changes be made to Analysis Plan?
(3) Is analysis really being done?
(4) Is analysis directed toward measures of effectiveness?
(5) Does analysis pertain to variation as well as to averages?
(6) Does analysis pertain to comparisons and relationships as well as to summarizations?
(7) Is analysis more than computer "push-button?"
(8) We are avoiding our #1 mistake-grouping horses and cows?
(9) Are we proceeding from detail to general?
(10) Are test conditions used actual rather than from Test Plan?
(11) In side-by-side comparison are we analyzing the differences rather than the data?
(12) What distribution is being assumed?

(13) Are data being over-analyzed?
(14) Are calculations being checked?
(15) Is analysis time being minimized?

c. MOMS (Measure of Mission Success)

(1) Is MOMS used?
(2) Is the MOMS meaningful?
(3) Does the MOMS answer project objectives?
(4) Has the MOMS procedure been verified?
(5) Have assumptions been noted?
(6) Have limitations been noted?
(7) Is the MOMS complete?
(8) We are including target evasion?
(9) Are we including the human element in the MOMS?
(10) Are we including "real" 1985 threats in our MOMS?
(11) Are all data sheets, worksheets, and tapes documented and filed?

d. Presentation of Results

(1) Are project objectives answered?
(2) Is each result given the "So What" test?
(3) Are we serving all of our readers?
(4) Are we presenting all we have learned?
(5) Are we reporting opinions as opinions?
(6) Are we presenting MOMS results first?
(7) Is language meaningful to all types of readers?
(8) Are we pointing out limitations, weaknesses, etc., in analysis procedures?
(9) Are data to be given in enclosure?
(10) Are sample sizes given in tables, graphs?
(11) Is measure of variation reported?
(12) Are confidence limits given?
(13) Is personal bias avoided?
(14) Is degree of coverage for each topic appropriate to importance?
(15) We are not exact (33.33%) in non-exact areas?
(16) Do we interpret tables and graphs for the reader?
(17) We don't imply differences when none exist?
(18) We don't give overall summary figures when they are meaningless?
(19) We don't confuse data with results?
(20) We don't permit Limitations to Scope to whitewash our report?

Section 2

Test Plan Rehearsal

201. Purpose. This is a procedure for measuring the efficiency of the testing and data analysis plan. The rehearsal is a paper and pencil exercise; however, use of computers or simulations may be appropriate. The rehearsal is "conducting" project operations (while still in the planning stage) at your desk.

a. Step 1: The Estimates. The following values must be estimated:

(1) The "average" expected values of the response. The values to be estimated are the values that would be expected if the response were measured at each of the test conditions specified in the Test Plan. Keep in mind that only "reasonable estimates" are needed. Where strong interactions are probable, the change in effect of one variable with the setting of the other should be considered.

(2) The variation expected in the response measurements. This can probably be derived from previous OPTEVFOR reports. Note again that only a reasonable estimate is required.

b. Step 2: Generating Fictitious Data. With these values, the next step is to generate the fictitious data. The procedure is quite simple. Add an error term to each of the estimated response values from step 1. A different value for the error term will be added to each response value. The different values to be added are derived using the variation estimated in step 2 and a table of random normal deviates. See Table 2-1. The fictitious data consist of a generated response (i.e., estimated value plus error) for each set of conditions at which a measurement of the response will be made during project operations.

c. Step 3: Analysis. The Project Analyst analyzes this fictitious data using the analysis plan he has decided to use for the actual data. This is done precisely the same way the actual data will later be analyzed. Conclusions on the fictitious performance should be drawn on the basis of the analysis.

d. Step 4: Measuring Effectiveness of the Test and Analysis Plans. One measure of the validity of the test and analysis plans is the difference between the performance predicted as a result of the fictitious analysis and that estimated in step 1. To determine this the Project Analyst should make a prediction of the response performance at each of the test conditions based on the results of the fictitious analysis. He then compares these predictions with the estimates of step 1. Since the data are based on these estimates and the expected experimental error, the

Table 2-1

Table of Random Normal Deviates

-1.40	-0.43	+0.22	-1.64	+1.80	+0.70	-1.04	+0.32	+0.63	+0.32
-0.36	-1.54	-1.41	+0.63	+1.43	-2.62	-0.12	-0.27	-0.13	+0.79
+1.30	-0.79	+2.55	+1.06	-0.49	+0.20	+1.29	-0.52	-0.23	+0.03
-1.38	-0.28	-0.53	+0.13	-1.45	-1.18	+0.97	+0.90	-0.00	+1.12
-1.40	+0.78	+0.67	-1.08	+1.14	+0.84	-0.29	+0.40	-0.43	+1.16
-0.56	-0.42	+1.21	-1.44	-0.55	+0.55	+1.39	+0.53	+0.54	-0.90
-0.47	+0.83	+0.38	+1.89	-0.50	-1.60	+0.04	-2.29	+0.89	-0.29
+1.38	+0.32	-0.48	-0.80	-1.33	+0.83	-0.67	-1.99	-1.12	+1.79
+0.00	+1.78	+0.18	+0.61	-1.18	-0.49	-1.08	+1.48	-1.05	+0.81
+0.63	-0.58	+0.80	-0.72	-2.02	+1.04	+0.06	+0.90	+0.01	-0.94
+0.57	+1.71	+0.72	-0.71	-1.75	+0.09	+1.00	+1.52	+1.00	-0.84
+2.01	+2.35	-0.41	-0.35	-1.30	+2.25	+0.28	-1.64	-0.22	-1.36
+0.49	-0.05	+1.32	-0.39	+0.08	-1.03	-1.84	-0.48	+1.69	-2.36
+1.29	+0.67	-1.33	-0.07	+0.58	-1.10	+0.28	-0.23	+0.39	-0.56
-1.13	+0.24	+0.65	-0.27	-0.14	-1.11	-0.16	+0.01	-0.44	-0.60
+0.55	+0.17	-0.65	+0.25	-0.42	-0.81	+0.88	-0.92	+0.28	-1.19
+0.90	-0.19	-1.70	+0.19	-0.61	-0.50	-1.52	-0.35	+1.32	+0.88
-0.75	+1.36	-0.85	+0.92	-1.91	-0.13	+1.98	+0.06	-0.62	-0.75
+0.80	-0.50	+0.41	+1.32	-0.07	-1.10	+0.66	+1.48	-2.42	+0.94
+1.21	-2.28	-0.29	-1.26	-0.12	+0.57	+0.84	-1.02	+1.67	+0.46
+0.23	+0.69	-1.32	-1.58	-0.31	+0.17	-0.63	+1.05	+0.90	+0.22
-0.33	-1.10	-0.97	-0.03	+1.61	-1.70	-1.70	+0.46	-0.92	+0.07
+0.77	+0.22	+0.67	+0.65	-1.46	+0.77	-1.02	-0.23	+0.57	-1.07
+0.03	+0.32	+0.73	-0.27	+1.54	-0.84	+0.75	+0.34	-0.38	+0.67
+0.40	-0.42	-0.43	-1.86	+1.11	-0.67	+0.29	-0.03	-0.23	-0.99
+1.33	-0.16	+0.62	+0.79	-0.78	+0.52	-0.44	+0.83	-1.12	-0.71
-1.15	-1.83	-0.56	-1.02	+0.02	+1.31	-0.82	-0.18	+0.59	+0.04
-1.30	-1.15	+1.98	+0.96	-1.93	+1.14	-0.73	-1.73	+1.86	-0.44
+0.50	-1.44	+0.32	-1.24	-0.11	-0.34	+1.31	-0.16	-1.03	-0.21
+1.49	-1.39	-0.30	+0.25	+0.60	-0.76	-0.74	+0.98	-1.47	-0.62
-0.43	+0.46	+0.06	+0.08	-3.13	+1.45	+0.28	-0.44	+0.59	+0.89
-0.03	-2.12	+0.07	+1.44	-1.27	-0.27	-0.78	+0.04	-2.46	-1.96
-1.15	+0.13	-0.28	-0.10	+0.21	-0.26	+0.03	+0.50	-0.99	-0.11
+0.03	-0.19	+0.08	+0.57	+0.85	+0.74	+0.78	+1.91	+0.43	+0.29
-1.18	-0.40	+0.34	+0.59	+1.27	+0.59	-0.07	+0.41	+0.43	-0.64
-0.30	-2.05	-0.70	-0.79	+0.44	-1.22	-0.64	+0.04	+0.85	+0.62
+1.22	-1.23	-0.79	+0.58	+0.73	+0.86	+0.64	+0.65	+1.10	+0.96
-1.37	-0.10	-0.74	+0.65	-0.90	+0.03	-1.56	+0.07	+0.33	-1.75
-0.97	+0.17	+1.95	+0.53	-0.16	+1.65	-2.71	+0.47	-0.24	-0.12
-0.87	+0.06	+1.34	+0.09	+0.14	-0.94	-0.44	-0.04	+2.90	+1.44
+0.44	+0.33	-1.09	+0.28	-0.82	-0.39	+0.48	-0.67	+1.21	+0.05
-2.34	-1.15	-2.27	-1.19	-0.50	-1.00	-0.97	-0.45	+0.95	+0.28
+2.54	-1.07	+0.14	-0.00	-1.00	+0.60	-0.08	-0.69	-0.46	-1.64
+0.08	+1.01	-0.15	+1.49	+0.09	-1.28	+0.25	+0.81	-0.12	+0.41
-0.69	-0.90	-0.16	-0.65	-1.20	+0.25	-1.07	-1.65	-1.85	-0.04
-0.56	-1.96	-1.41	+0.03	-1.20	-0.11	-0.91	-1.15	+0.62	-0.58
+0.66	-0.72	+1.78	-0.21	+2.50	+0.04	+0.81	+2.18	+0.91	+0.13
+0.74	-0.12	-0.74	+0.11	+0.18	+0.16	-0.28	+0.23	-1.46	+0.52
-0.81	+1.05	+2.61	-0.04	-0.17	-0.24	+0.26	-0.72	+0.91	+0.99
-0.55	-0.35	-0.82	+0.01	+0.45	+0.56	-0.30	+1.98	-0.43	-0.22

prediction should be close enough to the estimate to satisfy the Project Analyst that the project objectives will be met.

202. An Example of a Paper Rehearsal. Suppose one phase of the Test Plan calls for acquisition times to be measured at these eight test conditions:

Mode	Orientation	Speed (knots) 5	10
A	Best	1	2
	Worst	3	4
B	Best	5	6
	Worst	7	8

Suppose the plan also includes making these runs in two ocean areas. The conditions are to be tested only once at each location. The sequence of testing the eight conditions at each location is not specified in the plan. The plan is for the 16 valid measurements to be obtained even if invalid runs have to be repeated. The analysis plan for this phase is given in the Test Plan simply as, "An analysis of variance of the acquisition time data."

a. The Inputs

(1) Guesstimates are made of the acquisition times that are expected at the eight test conditions. These values are those considered to represent an inherent capability averaged over a wide variety of situations. Suppose the estimates are:

Mode	Orientation	Speed (knots) 5	10
A	Best	(1) 115	(2) 118
	Worst	(3) 115	(4) 118
B	Best	(5) 118	(6) 121
	Worst	(7) 122	(8) 125

(2) Normal fluctuations of the data about any one of these average values is to be expected. Suppose a maximum spread of $\pm$ 12 seconds is considered likely. This total spread, 24 seconds, is used to obtain an estimate of the standard deviation. (The standard deviation might be estimated from the results of similar projects conducted in the past. However, in this example a different method is shown for estimating the standard deviation.) The 24 seconds is the estimated range of the data. A rough rule of thumb, sufficiently accurate for the paper rehearsal, is -- divide the estimated range by 5 and use the result as an estimate

of the standard deviation. In this case 24/5 is 4.8; this is rounded to 5. The standard deviation is estimated to be 5 seconds.

(3) While the project plan calls for two locations, suppose this is not felt important with respect to this phase. That is, acquisition times are not expected to vary by location. However, to see if a large effect could be detected by this plan with the planned number of firings, the first location is assigned a +5 second value and the second location a -5 second value, i.e., a 10-second difference is built in for the rehearsal. To keep this illustration simple, no other environmental effects will be considered.

Note: The OTD is to supply these guesstimates. In order to do this, the Mode and Orientation variables have to be carefully defined. The Test Plan should be revised, if necessary, to reflect this careful definition. The expected values for Mode B might be obtained from a CNA simulation study. The values for Mode A may be based on judgment. It should also be noted that exact values are not required.

b. The Data

(1) Suppose the OTD considers that there is no real reason not to randomize the sequence firings. Then the eight conditions at each location can be randomized as follows:

Location	Test Sequence
First	3,6,5,8,4,2,1,7
Second	3,8,5,1,4,6,7,2

The acquisition times are obtained as illustrated in the Data Formation Worksheet, Table 2-2.

Table 2-2

Data Formation Worksheet

Sequence	Condition	Location	Normal Deviates	Values (sec) Conditions	Location	Error	Data
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	3	1	.05	115	+5	0.2	120.2
2	6	1	-.52	121	+5	-2.6	123.4
3	5	1	-1.41	118	+5	-7.0	116.0
4	8	1	1.82	125	+5	9.1	139.1
5	4	1	1.35	118	+5	6.8	129.8
6	2	1	.42	118	+5	2.1	125.1
7	1	1	-1.76	115	+5	-8.8	111.2
8	7	1	-.96	122	+5	-4.8	122.2
9	3	2	.56	115	-5	2.8	112.8
10	8	2	-.72	125	-5	-3.6	116.4
11	5	2	-2.10	118	-5	-10.5	102.5
12	1	2	.44	115	-5	2.2	112.2
13	4	2	.48	118	-5	2.4	115.4
14	6	2	-2.22	121	-5	-11.1	104.9
15	7	2	1.72	122	-5	8.6	125.6
16	2	2	-.94	118	-5	-4.7	108.3

Note: Col (4) is taken from Table 2-1.

Col 8 is derived by summing Col 5, Col 6, and Col 7. This latter value is the product of the standard deviation input (5 seconds) and Col 4.

(2) The derived data can be listed in the following form:

Mode	Orientation	Location: First Speed 5	First Speed 10	Second Speed 5	Second Speed 10
A	Best	111.2	125.1	112.2	108.3
	Worst	120.2	129.8	112.8	115.4
B	Best	116.0	123.4	102.5	104.9
	Worst	122.2	139.1	125.6	116.4

c. The Analysis. The analysis of variance is used in analyzing the data as specified in the Test Plan. Note: The actual clerical operations involved are given in Section 4, which uses the above data to illustrate the analysis of variance technique.

d. The Results

(1) The acquisition time (overall average 118 seconds) does not vary significantly by Mode, but does vary by Orientation and Location.

(2) The 2-second improvement in Mode A is not detected as significant.

(3) The average acquisition time for the best orientation is 113 seconds, while for the worst it is 123 seconds. This 10-second differential is significant.

(4) There is a 10-second improvement in acquisition performance for runs made at the second location as opposed to the first location.

(5) Speed is not important.

(6) No particular inconsistencies or interactions are detected.

(7) The experimental error standard deviation is 6.6 seconds.

The above results should be summarized as follows:

		Location			
		First		Second	
		Speed		Speed	
Orientation	Mode	5	10	5	10
Best	A				
	B	119		108	
Worst	A				
	B	128		117	

e. The Evaluation

(1) The process of running through paper rehearsal provides a check on test design while there is still time to modify it. For example, the rehearsal may show that a mass of measurements must be processed; this may lead to decisions to use punch cards, computer program preparation, etc.

(2) The primary purpose of the paper rehearsal is to check the Test Plan. So the final phase begins with comparing the results of the rehearsal with the known inputs. For example, Orientation is determined to be significant while Mode is not. Yet the inputs assigned a 5-second difference in the two modes

and only a 2-second difference in the two orientations. (An examination of Table 2-1 indicates the strong distortion because of the random play of the experimental error.)

(3) In this particular example, the test sensitivity may well be questioned since the plan could not detect the differences expected. If these differences are important and if the number of runs cannot be increased, then all phases of the plan should be reconsidered. Perhaps the approach should be changed to a sequential testing approach. Or perhaps an improved instrumentation system for determining acquisition times should be used. Experimental error may thus be reduced sufficiently to be worth it. The taking of other data to be used for "standardization" should be examined. There are many other considerations that may help. The paper rehearsal has given added direction and impetus to a reconsideration.

(4) An important benefit of the paper rehearsal is not illustrated in this example. This benefit is in comparing two different Test Plans. This is perhaps the most valuable use of the rehearsal.

Section 3

Simple Sum Squares Calculations

301. Arithmetic

a. For combining properties of the data X, the quantities obtained in arriving at the standard deviations are so important as to warrant special names. $\Sigma(X - \bar{X})^2$ are called Sum of Squares or SS for short. (These are actually sum of squares of deviations from the mean.) The square of the standard deviation is called the variance or s^2.

b. Calculation of SS involves finding the mean, rounding the mean, returning to the data and finding deviations from the mean. In practice, an identity is used to avoid these difficulties.

$$\Sigma(X - \bar{X})^2 = \Sigma(X)^2 - \bar{X}\,\Sigma X$$

The identity is used for calculating the sum of squares and standard deviation as follows:

Data (N = 4)

	X
	4.3
	6.2
	1.8
	3.9
ΣX =	16.2
$\bar{X}$ =	4.05
ΣX^2 =	75.38
$-\bar{X}\,\Sigma X$ =	65.61
$\Sigma(X-\bar{X})^2$ =	9.77
s^2 =	3.2567
s =	1.80

c. The term $\bar{X}$ (ΣX) is called the correction term for the mean. This term "corrects" the variation for the zone of the data. Thus, coding the data by translation does not affect the standard deviation, because this term will correct it. Change of scale, however, does affect the standard deviation.

302. Using Logarithmic Transformation (Base e)

Original Data (N = 4)	Ln Transformation	
4.3		1.459
6.2		1.825
1.8		0.588
3.9		1.361
	ΣX	5.233
	$\bar{X}$	1.308
	ΣX^2	7.657371
	$-\bar{X}\ \Sigma X$	6.844764
	$\Sigma(X - \bar{X})^2$	8.12607
	s^2	0.270869
	s	0.520

Note: There is no difficulty in finding the antilog of $\bar{X}$ and s. The mean is 3.7 units. The standard deviation is 1.68. This latter term must be interpreted further as 168%.

Note: We can use the multiplier factor (t) with the standard deviation tc form confidence limits. The best way to handle this is to form the product with s still in log form and then find the limits by finding antilogs. If t is 3, then ts = 1.560. This amount should be subtracted and added to the mean and then the antilogs found for the limits. The arithmetic gives us -0.252 and 2.861, which in antilogs gives 0.78 and 17.5 as limits.

Note: If we had not used the log transformation, the mean would be 4.05 and standard deviation would be 1.80. See paragraph 301. Using the same t factor (3.0) to find confidence limits as in above, we would arrive at a negative lower limit. With certain phenomena such as repair time, reaction time, detection range, etc., a negative value would be meaningless. This would show a definite need for transformation to logarithms.

Section 4

ANOVA (Analysis of Variance) Calculations

401. Balanced. This example illustrates a general procedure for obtaining the ANOVA measures. This general procedure pertains to a balanced situation only. The situation in this illustration is that described in Section 2 on paper rehearsal.

a. Data. The data for a 2^3 factorial, two replications, are as follows:

		Location			
		First Speed		Second Speed	
Mode	Orientation	5	10	5	10
A	Best	111.2	125.1	112.2	108.3
	Worst	120.2	129.8	112.8	115.4
B	Best	116.0	123.4	102.5	104.9
	Worst	122.2	139.1	125.6	116.4

b. Calculations. The steps in the calculations are:

(1) Firm up the model based on the Test Plan, actual runs, etc.

(2) Select a transformation of data if errors are deemed non-homogeneous.

(3) Obtain totals tables.

(4) Obtain ANOVA Sum Squares by applying the general procedure.

(5) Obtain ANOVA Mean Squares.

(6) Make F tests.

(7) Interpret.

c. The Model. The model is straightforward, being the sum of each of the variables and their interactions. The model also includes, as a measure of experimental error, a pooling of all interactions with the location factor. The model is the sum of the terms in the Effect column in the ANOVA worksheet, Table 4-2. The situation under analysis does not indicate any unusual mechanism involved. Hence, the data are not transformed.

d. Totals Table. The next step involves obtaining the various totals pertaining to the model. For example, the model includes a Mode by Speed interaction. This leads to a two-way table of mode and speed; the data at each mode/speed combination are totaled. The various tables could be obtained separately or at one time as is shown in the ANOVA Totals Table, Table 4-1. A scrutiny of this table will indicate that all of the totals needed are included. For example, the Mode by Speed totals have been obtained by totaling both Orientations, the totals being 456.4, 478.6, 466.3 and 483.8.

e. Sum of Squares. The SS for each term in the model is obtained by using the totals table in the following form.

$$SS = \frac{\Sigma T^2}{k} - c - SS_m \quad . \ . \ . \ . \ (1)$$

k is the number of data points going into each pertinent total. ΣT^2 represents a set of pertinent totals, each squared and the squares then summed. c is the correction (for the overall mean) term. c is simply the grand total of all data multiplied by the grand mean. SS_m are a series of Sum Squares by which the overall mean must be corrected. Thus, in the Mode by Speed measure, SS_m would be SS for Mode and SS for Speed. The above description will be clearer as the calculations are made.

(1) Find the correction term c.

$$c = \bar{X}\Sigma X = 117.8(1885.1) = 222{,}100.$$

(2) Find the Total Sum Squares. This is the sum of the squares of the individual data with c subtracted from this sum. The formula (1) is modified to (since $SS_m = 0$)

$$\text{Total SS} = \Sigma X^2 - c$$

$$\text{or } 111.2^2 + 120.2^2 + 116.0^2 + \ . \ . \ . \ + 116.4^2$$

$$- 222{,}100 = 1370$$

(3) Find the Mode SS. The Mode values from Table 4-1 are 935.0 and 950.1. These totals are each based on eight data points. Using formula (1) with $SS_m = 0$,

$$\text{Mode SS} = (935.0^2 + 950.1^2)/8 - 222{,}100 = 14$$

(4) Find the Orientation SS.

Orientation SS = $(903.6^2 + 981.5^2)/8 - 222{,}100 = 379$

(5) Find the Speed SS.

Speed SS = $(922.7^2 + 962.4^2)/8 - 222{,}100 = 98$

(6) Find the Mode by Orientation SS. The M by O table is 456.8, 478.2, 446.8, and 503.3. (Each of these totals is based on four data points.) However, the above four totals also have included (in their variation from each other) the effects of Mode and Orientation. So SS_m in Formula (1) is Mode SS and Orientation SS.

Mode by Orientation SS = $(456.8^2 + 478.2^2$

$+ 446.8^2 + 503.3^2)/4 - 222{,}100 - 14 - 379 = 78$

(7) The Mode by Speed SS is

$(456.4^2 + 478.6^2 + 466.3^2 + 483.8^2)/4 - 222{,}100 - 14$

$- 98 = 2$

(8) The Orientation by Speed SS is

$(441.9^2 + 480.8^2 + 461.7^2 + 500.7^2)/4 - 222{,}100 - 379$

$- 98 = 1$

(9) The M by O by S measure is

$(233.4^2 + 233.0^2 + 218.5^2 + 247.8^2 + 233.4^2 + 245.2^2$

$+ 228.3^2 + 255.5^2)/2 - 222{,}100 - 14 - 379 - 98 - 78$

$- 2 - 1 = 0$

(10) The Locations SS is obtained in a manner similar to a main effect SS.

Location SS = $(987.0^2 + 898.1^2)/8 - 222{,}100 = 493$

(11) The Remainder SS is the difference of all the SS measures from the Total SS.

Remainder SS = $1370 - 14 - 379 - 98 - 78 - 2 - 1 - 0$

$- 493 = 305$

f. Degrees of Freedom. The degrees of freedom is a measure of the effective sample size available to determine a particular

effect. Usually this is merely the number of settings minus one. For the Mode variable, there are two settings in the study, Mode A and Mode B. The degrees of freedom for this effect is 2 - 1 = 1. Note that the degrees of freedom for interaction terms are the product of the degrees of freedom for the respective variables present in the effect.

g. Mean Square. The Mean Square values are obtained by dividing the respective Sum Squares by the corresponding degrees of freedom.

h. F-Test

(1) To determine the importance or significance of an effect, its Mean Square is compared to that of the error term. Ratios less than one can be ignored, i.e., they imply no evidence of importance. Ratios larger than one are of interest. Tables are available in standard statistical texts that relate magnitudes of these ratios (and the corresponding degrees of freedom) to significance levels.

(2) The Orientation Mean Square is 379, which, when compared to the Remainder or error term, leads to a ratio of 8.6. Using the standard F tables, this ratio corresponds to a significance level of about 0.03, which means the probability is 0.03 that the observed difference in the Orientation settings could have occurred simply by chance. This information is useful in determining the decision to be taken, the conclusions to be made, etc. All of the terms in the model are tested in a similar manner.

Table 4-1

The ANOVA Totals Table

Sum of both locations

Mode	Orientation	Speed 5	Speed 10	Speed Both
A	Best	223.4	233.4	456.8
	Worst	233.0	245.2	478.2
	Both	456.4	478.6	935.0
B	Best	218.5	228.3	446.8
	Worst	247.8	255.5	503.3
	Both	466.3	483.8	950.1
Both	Best	441.9	461.7	903.6
	Worst	480.8	500.7	981.5
	Both	922.7	962.4	1885.1

First Location: 987.0
Second Location: 898.1

Table 4-2

The ANOVA Worksheet

Effect	Degrees of Freedom	Sum of Squares	Mean Square	F Test
Mode (M)	1	14	14	
Orientation (O)	1	379	379	8.6
Speed (S)	1	98	98	2.2
M by O	1	78	78	1.8
M by S	1	2	2	
O by S	1	1	1	
M by O by S	1	0	0	
Locations	1	493	493	11.2
Remainder	7	305	44	
Total	15	1370		

402. Unbalanced (One Variable)

a. The Sum of Squares calculations for ANOVA are straightforward for the one-variable case even when sample sizes vary by test levels. For example, reaction time test data (with log transformation):

	Scenario					
	S	M	L	s	m	l
	.792	.924	.477	1.097	1.000	1.340
	.708	.681	.462	1.004	1.127	.892
	.968	.903	.531		1.152	.924
			.633		1.072	1.009
						.903
Σt	2.468	2.508	2.103	2.101	4.351	5.068
$\bar{t}_2$	.823	.836	.526	1.051	1.088	1.014
Σt^2	2.065552	2.132946	1.123623	2.211425	4.746417	5.278530
c	2.030341	2.096688	1.105652	2.207100	4.732800	5.136925
SS	.035211	.036258	.017971	.004325	.013617	.141605

b. The ANOVA is as follows:

Effect	D/F	Sum Squares
Among Scenarios	5	0.836992
Within Scenarios	15	0.248987
Total	20	1.085979

c. The Sum of Squares for Among Scenarios is the sum of the products $\bar{t}\Sigma t$ or c for each scenario minus the grand c term for all the data.

d. The Sum of Squares for Within Scenarios is merely the sum of each scenario SS.

e. The Sum of Squares Total is the sum of the squares of all the data minus the grand c term.

f. The ANOVA would include the mean squares, etc., which are straightforward and not shown above.

403. Unbalanced (Two Variables)

a. The ANOVA with two-way classification is from Rao, Advanced Statistical Methods in Biometric Research. Note: One variable must be at two levels.

	B_1	B_2 . . . B_p
A_1	$\bar{X}_{11}$	$\bar{X}_{12}$ . . $\bar{X}_{1p}$
A_2	$\bar{X}_{21}$	$\bar{X}_{22}$ . . $\bar{X}_{2p}$
Difference (d_i)	d_1	d_2 . . . d_p

Weights (w_i)	$\frac{m_{11}\ m_{21}}{m_{11}+m_{21}}$	$\frac{m_{12}\ m_{22}}{m_{12}+m_{22}}$	$\frac{m_{1p}\ m_{2p}}{m_{1p}+m_{2p}}$
$w_i d_i$	$w_1 d_1$	$w_2 d_2$ . . .	$w_p d_p$
$w_i d_i^2$	$w_1 d_1^2$	$w_2 d_2^2$ . .	$w_p d_p^2$

For illustration the data in 402. will be used. The scenarios are related as follows:

Scenario Code

Target Size (S)

Environment (E)	Small	Medium	Large
Clear	S	M	L
Jam	s	m	l
	Sample Size		
Clear	3	3	4
Jam	2	4	5
Weights	1.20	1.71	2.22
	Averages		
Clear	.823	.836	.526
Jam	1.051	1.088	1.014
d	.228	.252	.488
wd	.2736	.4309	1.0034
wd^2	.0624	.1086	.5287

The target size/environment interaction SS is found first by

$$w_i d_i^2 - \frac{(w_i d_i)^2}{w_i} = .6997 - \frac{(1.7879)^2}{5.13} = .0768$$

Small	Medium	Large
.792	.924	.477
.708	.681	.462
.968	.903	.531
1.097	1.000	.633
1.004	1.127	1.340
	1.152	.892
	1.072	.924
		1.009
		.903

Target Size (Ignoring Environment	.137171
Interaction	.076580
Environment (By Subtraction)	.623241
Among Scenarios	.836992

In a similar manner:

Environment (Ignoring Target Size)	.6032919
Interaction	.076580
Target Size (By Subtraction)	.157120
Among Scenarios	.836992

b. The Sum Squares obtained by subtraction are valid and are used in the final ANOVA given as follows:

ANOVA

Environment	1	.623241
Size	2	.157120
Interaction	2	.076580
Among Scenarios	5	.836992*
Within Scenarios	15	0.248987
Total	20	1.085979

*Non-additive due to unbalance.

404. Unbalanced (Complex Case). For complex cases, the formal ANOVA structure is best reduced to multi-sets of single contrasts and multiple regression techniques are used. See the examples in paragraphs 504 and 2002.

Section 5

Regression

501. Fitting a Straight Line

a. During an evaluation of an aiming device, 31 retarded bombs were dropped from aircraft. The retarded device was somewhat unusual. And the drops were in a jungle environment. The bombing errors (E) along the range axis were the important data. The average observed E was -2 ft; E varied from -150 to +165 ft, with a standard deviation of 75 ft.

b. The wind varied widely during the evaluation. Wind data were available, but there was some question as to its value in the jungle atmosphere. To determine this, the wind component along the range axis (W) was found and plotted as in Figure 5-1. Note that this working plot denoted the pilot making each drop. The trend, if any, did not seem to be due to any particular pilot, nor did the trend seem to vary with pilots. The pilots were thereafter ignored in the analysis.

c. Figure 5-1 indicates that a trend, a simple straight line, may be present. It was decided to fit this line with regression analysis. The line to be fitted was

$$\text{Error (E)} = a + b\ \text{Wind (W)}$$

d. The data can be considered as 31 pairs (N = 31) of data points; one column of range errors (E) and a corresponding column of wind values (W). This straight-line fitting has been programmed on the Wang computer. As a matter of interest, the following sums and formulae apply:

$$\Sigma E,\quad \Sigma(E^2),\quad \Sigma W,\quad \Sigma(W^2),\quad \Sigma(WE)$$

The formulae to obtain the two unknowns, a and b, were:

$$(1)\ \ldots\ldots\ b = \frac{\Sigma we}{\Sigma w^2}$$

$$(2)\ \ldots\ldots\ a = \bar{E} - b\bar{W}$$

The right side of (1) and (2) were obtained as follows:

$$\Sigma we = \Sigma WE - \bar{W}(\Sigma E)$$

$$\Sigma w^2 = \Sigma(W^2) - \bar{W}(\Sigma W)$$

A bar indicates an arithmetic mean of the indicated values. Applying the above, b was found to be 6.4 and a was 9. The line was

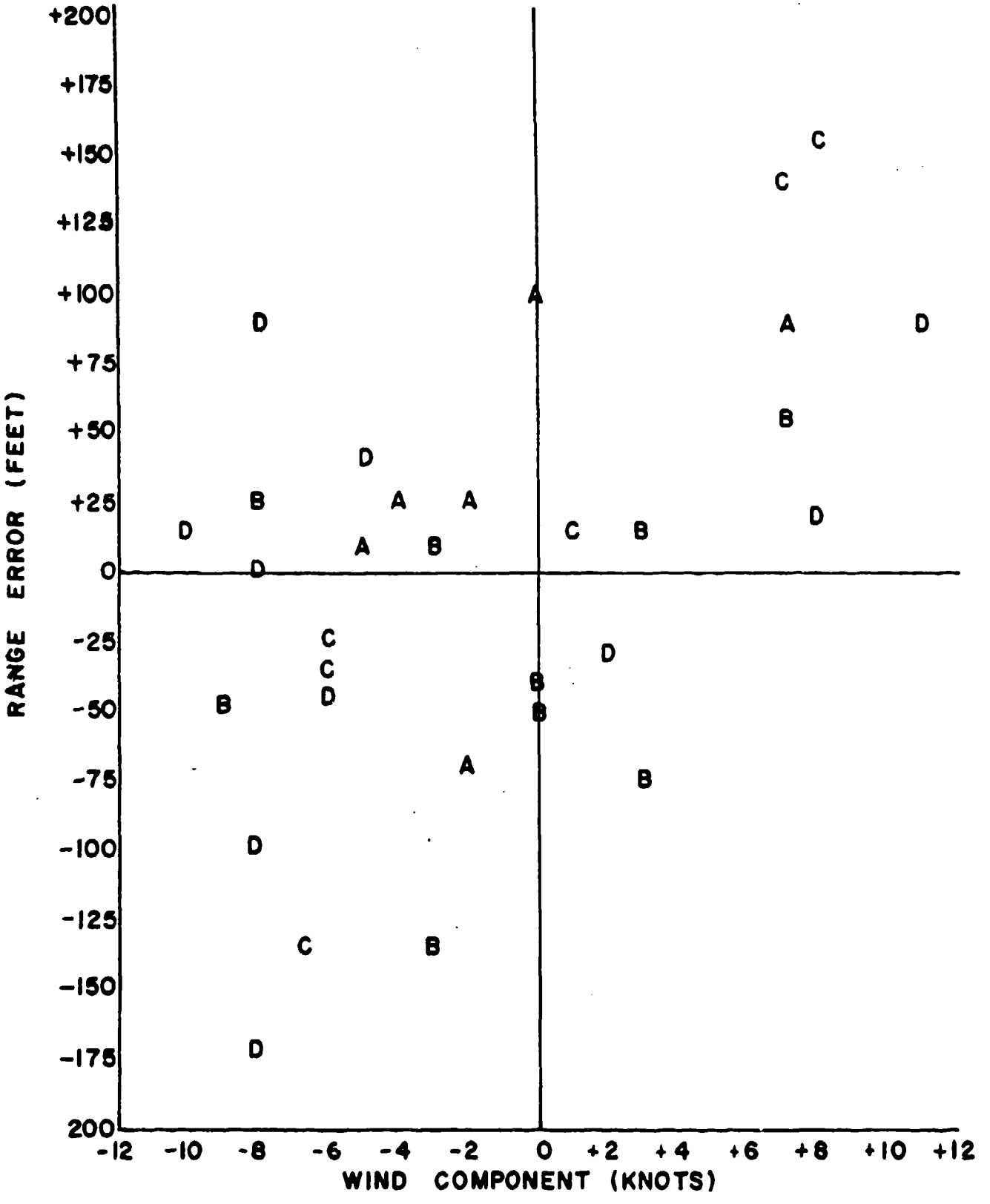


Figure 5-1

Scatter Plot: Range Error by Wind Component by Pilot

$$E = 9 + 6.4W$$

e. Before the equation was used, it was tested for significance. The correlation index r^2 may be found by

$$r^2 = \frac{b\Sigma we}{\Sigma E^2 - \bar{E}(\Sigma E)}$$

The fit was found to be significant; see Table 5-1; the correlation index (r^2) was 26%.

f. Figure 5-2 shows the fitted line.

g. The analysis operation continued with an attempt to include more than the wind variable. Altitude at release looked promising. The equation fitted was:

$$\text{Error} = a' + b' \text{ Wind} + c \text{ Altitude}$$

The parameters a', b' and c were found by an extension of the formulas given. However, significance testing indicated that altitude was not important in influencing errors.

502. Interpretation

a. The significant trend was found to be

$$E = 9 + 6.4W$$

When there was no wind present (W=0), E is 9 ft. This measure of accuracy can be useful in determining a need for recalibration, etc. Note that this value differs from the error average of -2 ft previously found which ignored wind. Thus the accuracy which should be reported has been changed.

b. The wind coefficient, b = 6.4, is interpreted as a ratio term; the range error is 6.4 ft per knot of wind along the range axis. One use of this measure is to standardize the data. For example, Pilot D had a 92-ft overshoot. The wind along the range axis was 11 knots. The adjustment factor for this is 6.4(11) = 70 ft. So if the situation was to be put in standard (zero wind) terms, the overshoot for this particular shot would be 22 ft (92 - 70). This procedure could be done for each shot. The pilot variable could then be restudied with adjusted data. This restudy would be more meaningful because it would compare pilots under the same (zero) wind condition.

c. The data adjusted as described above would also be useful in obtaining precision measures. For example, the adjusted standard deviation was found to be 66 ft, which is a significant reduction from the original 75 ft. In like manner the CEP value would be reduced. Thus, the precision has changed.

Table 5-1

The r^2 Table

Significance Level for r^2
By Degrees of Freedom (n = N - 2)
Simple Linear Only

n	Significance Level 0.10	0.05	r^2(%)	0.02	0.01
4	53	66		78	84
5	45	57		69	77
6	39	50		62	70
7	34	44		56	64
8	30	40		51	59
9	27	36		47	54
10	25	33		43	50
11	23	31		40	47
12	21	28		38	44
13	19	26		35	41
14	18	25		33	39
15	17	23		31	37
16	16	22		29	35
17	15	21		28	33
18	14	20		27	32
19	14	19		25	30
20	13	18		24	29
25	11	15		20	24
30	9	12		17	20
50	5	8		10	13
90	3	4		6	7

d. The correlation index (r^2) was found to be 26%. This means that 26% of the variation in range error could be explained by the wind. This measure of efficiency of fit indicates that wind is an important variable. However, it also indicates that wind is not the sole variable; there are other extraneous effects in the system that have not been taken into account.

e. As an aside, the bi-directional method of bomb evaluation is another method of wind adjustment. This method is to conduct our testing in pairs of drops at opposite headings. The errors of the pairs are averaged. The averages are taken as free of the wind effect. (This is valid if the intercept term (a = 9 ft in the example) is minor.)

f. As an aside, the straight line formulas can be used directly to fit various types of relationships by use of transformations. The most important types are (to either base 10 or e):

(1) $\log Y = b_o + bX$

(2) $Y = b_o = b \log X$

(3) $\log Y = \log b_o + b \log X$

For example, we may be dealing with insulation resistance, R, as a function of temperature, T. From our plotting or from a priori considerations, we may decide to fit

$$R = kT^b$$

By transforming our data to logs, $\log R = Y$ and $\log T = X$, we would proceed to fit

$$\log R = \log k + b \log T \text{ or}$$

$$Y = b_o + bX$$

The usual linear-fitting formulas apply. However, we must realize that the least squares we are minimizing are the logs of R, not R directly.

g. The regression analysis method of standardizing is not foolproof. It must be used carefully. When properly used, it can have an important influence on our evaluations.

503. Multiple Regression

a. In an evaluation of a weapon system, the torpedo component was considered independent of the other components. There were 21 successful firings that had a complete set of measure-

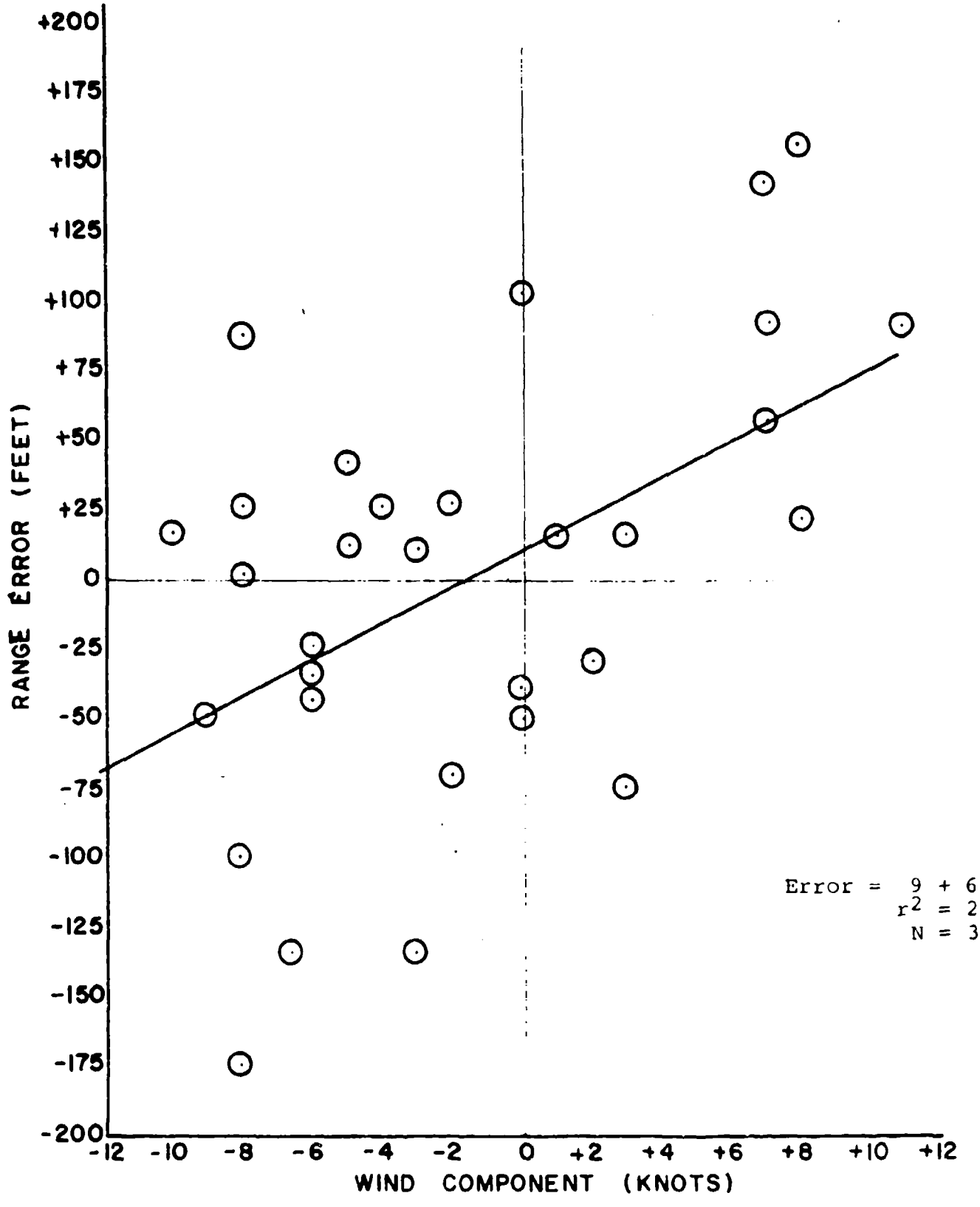


Figure 5-2

Line of Best Fit: Range Error by Wind Component

ments. Rather than simply satisfy ourselves with the fact that these 21 hit the target, we attempted to learn more by a study of time-to-hit as the critical parameter. The 21 firings had differing target speeds and target depth to torpedo search depth differentials (D). In addition, the weapon system had different total attack errors (TAE) for each weapon firing. This was uncontrolled but measured.

b. The equation of relationship fitted was:

$$\text{Time} = a' + b' (\text{TAE}) + c' (\text{Speed}) + d' (\text{Depth})$$

The speed variable was found not significant. Then

$$\text{Time} = a + b (\text{TAE}) + d(D)$$

was fitted. These two variables were found to be significant in their influence on time-to-hit.

c. The correlation index (r^2) was found to be 58%. This magnitude indicates that other variables also influenced time-to-hit. If the data were standardized, this magnitude would indicate what the variation would be in the adjusted time-to-hit data. The ratio of the adjusted to the original standard deviations would be 1-0.58 or 0.65.

d. The values of a, b, and d were next interpreted. The value of a is the minimum time-to-hit. Even if TAE and D were zero, there would still be an a (dead time). It turned out that this value was exactly the same (to the nearest second) as the rated time for the torpedo to reach the initial search depths (averaged). The coefficient b for TAE was very close to the rated speed of the torpedo. However, the d value for D was an order of magniture different than the rated depth rate.

e. While the above information could have been further explored, the real purpose of the analysis was to study the firings that failed to hit. The question considered, was "did these fail to hit because of large TAE and depth differential values?" The time-to-hit values were calculated using the above equation and the TAE and D for the firings that did not hit. None of these time-to-hit values approached the maximum designed time of running specifications. Thus, there was no evidence that the TAE and/or Depth differentials were important in causing misses.

f. As an aside, the multiple regression procedures and curve-fitting procedures can be interchanged. This is done by a direct transformation. For illustration, suppose in 501 we wished to fit not a linear in wind but a quadratic. This is fitting:

$$E = a'' + b'' w + b'' w^2$$

The multiple regression formulae can be used directly by transforming w^2 into a "multiple" variable.

g. In fitting polynomials when the independent variable is equally-spaced, orthogonal polynomials can be used to avoid matrix inversion. See paragraph 603.

504. Step-Wise Regression, or Data Analysis in the Unbalanced Situation

a. Introduction. In our evaluations, the data available for analysis are usually unbalanced as to test conditions. Table 5-2 illustrates a typical situation. We have always had extreme difficulty in analysis in such cases. With the advent of the computer, workers in other fields have modified a standard analysis technique for such situations. The technique is multiple regression, an extension of simple curve-fitting to a multi-variable case. The modification is the use of "individual degrees of freedom" rather than fitting for a trend. Most computers already have programs available for applying the multiple regression technique. For the unbalanced case, this technique is considered the best, the most valid, and the most accurate when used correctly. This paper demonstrates the technique using Table 5-2 data.

b. The Procedure

Step 1: The levels of the variables in Table 5-2 were put in quantitative form by coding as described in Table 5-3.

Step 2: The model to be fitted was set up. In this case, the model included all single variables and many two-variable interactions. This was limited by the computer program, which happened to handle only 26 terms in the model. The specific model used is given in Table 5-4.

Step 3: Step-Wise Regression program was used.

Step 4: The computer output indicated that only four terms in the model were significant. More important, the technique showed how these terms affected the model. The fitted coefficients are given in Table 5-5.

Step 5: These coefficient were used with the coding in Step 1 to translate the computer output back to the test conditions.

Step 6: The results were put in summary form to facilitate technical interpretation. See Table 5-5. This summary form would also aid in preparation for report presentation, i.e., probability of detection curves could be formed, etc.

Table 5-2
Detection Range by Test Conditions

Target Size and Profile	Altitude, Readiness and Weather									
	Very Low			Low		Medium		High	
	III Good	III Bad	I Good	I Bad	IA Good	IA Bad	I Good	I Bad	IA Good	IA Bad	I Good	I Bad	IA Good	IA Bad
Small Incoming	683	480	586	552	509	701	1114	779	848*	1211	1546	1332	1233*	1453*	1307*	1043*	1498	1438*
	517	342	628	519	590	782	1180		864*		1418*	1252*	1270*				1218	1482*
	805		639		541	745	1027				1385		1485				1362*	
					614		963*	735*	965*		1400		1354*				1500	
					674			720*	821*		1509*		1407				1514	
					516*												1538*	
					512													
Small Crossing		412	587		430*	746					1394	1013*	1203	1482	1424		1473	
		426			511*	670							1766	1608	1286*		1553	
					422*												1232*	
					522*					680							1454	
Large Incoming			603		512	588				1147								
			489		548													
					800*													
Large Crossing			419	398	609	682									1349*	1038*		
			600		327										1447			
					635													
					611													
					482													
					416													

*Denotes a radar mode change

Table 5-3

The Coding Used

Altitude	A1	A2	A3
Low	-1	+1	+1
Medium Low	-1	-1	-1
Medium High	+1	-1	+1
High	+1	+1	-1

Readiness Condition	C1	C2
III	-1	-1
I	0	+2
IA	+1	-1

Weather	W	Size	S
Good	+1	Small	-1
Bad	-1	Large	+1

Profile	P	Radar Mode	R
Incoming	+1	A	+1
Crossing	-1	B	-1

Note: The actual values for altitude would have been used if the curve-form was known.

Table 5-4

The Specific Model Used

Detection Range = a + b A1 + c A2 + d A3 + e C1 + f C2 + g W + h S + i P + j R + k A1 C2 + l A1 W + m A1 S + n A1 P + o A1 R + p C2 W + q C2 S + u C2 P + v C2 R + w W S + x W P + y W R + z S P + aa S R + bb P R + cc A2 C2

Notes: (1) To the limit of the computer program the model duplicates a typical analysis of variance model including all main effects and most two-variable interaction.

(2) The unknown coefficients are to be determined by the computer program using multiple regression. This technique is described in most statistical texts.

Table 5-5

Summary of Procedure Output

Detection Range = $1034 + 317A_1 - 104A_3 + 90C_1 + 35\ C_2W - 70\ SR$

The standard deviation after the completion of the procedure was 113.

No other effects were found significant.

The translation of the above equation into more meaningful terms gives the following:

ALTITUDE	RESULT
Low	613
Medium Low	821
Medium High	1247
High	1455

CONDITION	RESULT
III	-90
I	0
IA	+90

	CONDITIONS		
WEATHER	III	I	IA
Good	-35	+70	-35
Bad	+35	-70	+35

	SIZE	
RADAR Mode	Small	Large
A	+70	-70
B	-70	+70

The above detection range averages or the detection range model directly can be translated to different scenario situations.

c. Caveat. The accuracy of the model is, of course, critical. As with any technique, if the model has shortcomings, so will the results. This procedure is feasible only with use of a computer. Careful editing of data becomes more critical as we lose the tender loving care given to the individual data if we had used the hand calculator. Careful preparation for the computer program and careful interpretation of results are critical.

Section 6

Precision Determination and Accuracy

601. Grubbs's Method

a. Introduction. In accuracy trials of localization equipment, such as range and bearing accuracy tests of sonar, we use the equipment in conjunction with a reference or standard, e.g., an optical instrument. The accuracy of the equipment under evaluation can never be obtained independent of the accuracy of the standard. Likewise the precision of the equipment under evaluation may be influenced by the precision of the standard. If the precision of the standard is relatively very good, then there is no difficulty. If the precision of the standard approaches that of the equipment under test, the analysis must take this into account. This example illustrates an analysis technique that determines the precision of each of the measuring systems used. This technique determines the precision of the test equipment independently of the precision of the reference system used. This is useful when we are concerned with the possible poor precision of the reference system. This example also discusses the difference between precision and accuracy.

b. The Procedure

(1) An artificial set of data was generated to illustrate the procedure. Forty runs were made. Each run was made under different conditions. The data varied from 36 to 112. Three measuring systems (A, B, and C) were used. These represent, for example, an acoustic system under evaluation, Loran C, and DRT. In this artificial set of data, system A had a built-in precision of 2.8; system B had a precision of 1.0; system C had a precision of 4.6; system B is the most precise and system C is least precise. The actual measurements obtained are not reported here since they are not needed to illustrate the method. Table 6-1 gives the individual differences of the measurements obtained. Table 6-2 is a worksheet that gives the formulae to be used. As derived in the worksheet, the analysis procedure calculates the respective precision measures as 2.9, 1.6, and 4.7. This is good agreement with the 2.8, 1.0 and 4.6 built in this artificial set.

(2) The analysis technique is from a paper by Dr. Frank E. Grubbs, (JASA, June 1948). See also: "Two Simultaneous Measurement Procedures: A Bayesian Approach" by Draper and Guttman in JASA, Mar 1975. See also Jaech's paper in Technometrics, May 1976 and also Grubbs paper in Technometrics, Feb 1973. For a three instrument case the technique involves the various differences A-B, A-C, B-C for each run and then finding the standard deviations of these differences from which the individual system standard deviations are found. The technique

Table 6-1

Data Differences (40 Runs)

A-B	A-C	B-C	A-B	A-C	B-C
3.3	-1.5	-4.8	-1.6	-0.8	0.8
-0.3	-1.4	-1.1	0.1	2.8	2.7
-3.3	0.8	4.1	2.0	-5.8	-7.8
-4.3	-9.0	-4.7	-2.1	0	2.1
-1.0	-2.7	-1.7	4.4	8.0	3.6
2.8	1.6	-1.2	-4.3	-5.7	-1.4
0.8	5.4	4.6	-5.8	8.4	14.2
-0.7	-1.7	-1.0	-5.3	-5.0	0.3
0.1	-5.2	-5.3	6.8	-1.0	-7.8
0.4	5.2	4.8	0	-4.9	-4.9
-1.1	-5.9	-4.8	2.0	6.0	4.0
-3.3	-7.4	-4.1	0.9	-1.0	-1.9
-1.2	-8.9	-7.7	-1.8	-5.2	-3.4
5.9	-1.7	-7.6	1.3	4.2	2.9
4.0	1.8	-2.2	2.8	9.8	7.0
-2.3	-3.5	-1.2	-0.6	9.4	10.0
1.3	6.8	5.5	6.3	5.7	-0.6
-5.4	-6.5	-1.1	-2.3	-8.3	-6.0
-1.9	-2.8	-0.9	2.5	6.7	4.2
-6.6	-5.8	0.8	0.8	4.0	3.2

Table 6-2

Worksheet

A, B, C are the three measurement systems. (The data for each system are recorded but are not given due to space.) The symbol d, is the difference listed in Table 6-1, $\bar{d}$ represents the mean difference. n(=40) is the number of differences. S is the standard deviation. The usual formula is used as follows:

$$S^2 = \frac{\Sigma(d - \bar{d})^2}{n-1} = \frac{\Sigma d^2 - \bar{d}\Sigma d}{n-1}$$

When this is applied to each of the three columns in Table 6-1, we obtain

$$S^2_{A-B} = 10.86,\ S^2_{A-C} = 30.28,\ S^2_{B-C} = 24.68$$

Grubbs's paper gives the formulae below

$$S^2_A = 1/2\ (S^2_{A-B} + S^2_{A-C} - S^2_{B-C})$$

$$S^2_B = 1/2\ (S^2_{A-B} + S^2_{B-C} - S^2_{A-C})$$

$$S^2_C = 1/2\ (S^2_{B-C} + S^2_{A-C} - S^2_{A-B})$$

Substituting we get

$$S^2_A = 8.23,\quad S_A = 2.9$$

$$S^2_B = 2.63,\quad S_B = 1.6$$

$$S^2_C = 22.05,\quad S_C = 4.7$$

Note (1): The extension to more than three measuring systems is straightforward.

(2) As stated on Page 6-1, with two measuring systems, say A and B, the testing must be a series of repeat runs at the same, single condition. The formula then becomes:

$$S^2_A = \tfrac{1}{2}(S^2_X - S^2_Y + S^2_{X-Y})$$

$$S^2_B = \tfrac{1}{2}(S^2_Y - S^2_X + S^2_{X-Y})$$

where S^2_X and S^2_Y and S^2_{X-Y} are the variances of the original data of A and B and differences correspondingly.

needs three (or more) measurement systems with independent readings taken simultaneously on each system. If we have only two measuring systems, the runs must be repeat runs (at the same condition). The procedure also assumes that the precision does not change during the runs. In some projects this may not be the case, the precision may be relative to the magnitude being measured. In such cases we may have to analyze the data in relative (logarithmic) terms. In other cases we may be forced to divide the set of data into groups of equal magnitude of response. Each group would then be analyzed.

c. Precision and Accuracy

(1) Precision is defined as agreement of the data among themselves. Accuracy is defined as agreement of the data to a "true" value. In a sonar range accuracy test we may take a series of repeat readings. Suppose the data were 89.2, 89.1, 89.3, 89.2, and 89.2. We would say that this sonar was very precise; readings were close together. From these data alone though, we cannot comment on the accuracy. A reference system or knowledge as to a "true value" must be brought into play to determine accuracy.

(2) Although we usually consider precision and accuracy as related, they are determined by two different parameters and are actually unrelated. The accuracy property in the above set of data would be determined by the average of 89.2. If the "true" value was actually 89.2, then the system would be accurate as well as precise. If the "true" value was 98.2, then the difference of this from the mean of 89.2 would indicate the inaccuracy or bias in the system. The system would still be precise, but it would also be inaccurate.

(3) In the set of data used in Table 6-1, no bias was built into the system. The averages of the differences were minor and we would conclude all three systems were similar in accuracy. However, if the average of one system had been found to differ widely and significantly from another, then we would conclude that the accuracy was not similar. We would have need for a reference to determine which is more accurate.

602. Variate Difference Method

a. T.C. Cantwell, in a Hughes Aircraft Co memo of 28 May 1971, describes a method of precision/accuracy determination with one target and measurement system. Reasonably smooth target paths are assumed. The data measurements are equally spaced in time.

b. The variate difference analysis technique makes the implicit assumption that the actual target trajectory (range or azimuth versus time) may be fitted by a low order polynomial (although

this polynomial is never found). By taking successive differences in measured data (analogous to differentiation of a continuous function) the actual trajectory may effectively be eliminated from the measured data, with the resulting variations due only to measurement errors. For example, if the actual (but unknown) trajectory of target range versus time was of second order

$$r(t) = at^2 + bt + c$$

then the first difference would reduce this component of the measured data to a first order curve

$$\frac{d}{dt} r(t) = 2at + b$$

where the differentiation is analogous to differencing. The second difference would result in the constant

$$\frac{d^2}{dt^2} r(t) = 2a$$

and the third difference would eliminate the original trajectory completely

$$\frac{d^3}{dt^3} r(t) = 0$$

Since variances are calculated as deviations about the mean value, and after the second difference the trajectory becomes a constant (no variation with time), variances calculated from the second and all higher differences would be due only to measurement errors and would therefore allow range measurement error to be estimated.

c. For actual target trajectories, the contributions to the variance caused by the trajectory diminish with higher differences, but may never disappear completely. However, for most aircraft paths, the contribution due to the trajectory becomes negligible after two or three differences, becoming several orders of magnitude below the variance caused by measurement errors.

d. As an example of the variate difference analysis technique, assume the data of Table 6-3 represent the range of a target of interest on 10 successive radar scans. A calculation of the variance of these data results in 33.7. Table 6-4 shows the results of successive differences of the original data and the variance resulting in each case (after normalization; see Table 6-5). As can be seen, the variance resulting after differencing is relatively constant over the differences evaluated and is about 0.4. Since the original data were generated by perturbing a linear

Table 6-3

SCAN	Range
1	49.8
2	52.3
3	53.4
4	56.7
5	57.7
6	59.6
7	62.8
8	63.7
9	65.5
10	68.5

Table 6-4

DIFFERENCE

SCAN		0	1	2	3
1		49.8	-	-	-
2		52.3	2.5	-	-
3		53.4	1.1	-1.4	-
4		56.7	3.3	2.2	3.6
5		57.7	1.0	-2.3	-4.5
6		59.6	1.9	0.9	3.2
7		62.8	3.2	1.3	0.4
8		63.7	0.9	-2.3	-3.6
9		65.5	1.8	0.9	3.2
10		68.5	3.0	1.2	0.3
Var	=	33.7	0.4	0.5	0.5

Table 6-5

Normalization Values for each Difference Order

Difference n	Normalization C_n
1	2
2	6
3	20
4	70
5	252
6	924

Note: The estimate of the measurement variance obtained for the n^{th} difference is

$$V(n) = \frac{\text{Var}\,[\Delta_i n]}{C_n}$$

where C_n is a normalization constant given by

$$C_n = \sum_{j=0}^{n} \binom{n}{j}^2$$

where

$$\binom{n}{j} \triangleq \frac{n!}{j!(n-j)!}$$

function with noise having a variance of 0.25, the method is seen to derive the noise (measurement) variance quickly and easily (within sample size constraints) without actually determining the underlying trajectory.

e. This section has only briefly presented the variate difference technique. The general technique is as follows:

(1) The variance of the measured values is calculated. Variance = SS/n, not SS/n-1.

(2) The first difference of the measured values is calculated and the variance of the resulting values determined. Use Table 6-5 for normalization value.

(3) The two variances are compared. If they are sufficiently close together, the original trajectory is assumed to have been eliminated and the resulting variance is the measurement error.

(4) If the two variances are not sufficiently close together, the second difference is calculated and the variance is determined.

(5) The variances resulting from the first and second differences are compared. If they are sufficiently close together, the original trajectory is assumed to have been eliminated and the resulting variance is an estimate of the measurement error.

(6) If the two variances are not close together, the next difference is calculated, a new variance is determined, and new comparisons are made.

(7) The process repeats until variance convergence occurs (the resulting variance is an estimate of the measurement error) or until such a high order difference is reached that it is assumed that data cannot reasonably be represented by a smooth polynomial. In this latter case, the data are discarded or are partitioned into (hopefully smoother) subsets and the process is repeated on each subset.

f. There exists some difficulty in defining how high an order of differencing may be allowed before the decision is made that an adequate fit cannot be obtained. This would seem to be a matter of judgement, since a rigorous analysis cannot be made as to the goodness of fit of a polynomial to all possible target trajectories. However, some qualitative bounds may be set. There is the trivial bound that no more differences may be taken than the number of measured data points, since each difference removes one point from the data set. A more realistic bound may

be obtained by analyzing the order of curve required to fit typical target trajectories. From these analyses it appears that about four differences are adequate to fit most smooth trajectories. At this time it appears that no more than four differences should be taken, and that the initial data set should be large enough that four differences will not substantially degrade the sample size (e.g., at least 30 samples).

603. Orthogonal Polynomials

a. The variate difference method, previously described, is useful when the data measurements are equally spaced, say in time. With the assumption of equally-spaced data, a standard curve fitting method, called orthogonal polynomials is available. This method determines the actual polynomial which "fits" the data best (in the statistical sense). The variation remaining in the data after fitting is taken as a measure of precision of the measuring system. The orthogonal polynomial method is described in many statistical tests such as reference (a).

b. The target range data of 602 above was fitted by the orthogonal polynomial procedure. Polynomials higher than linear were found to be not significant. The best fit was then a linear. Range = 47.92 + 2.0 Scan. The correlation index, r^2 was 99%. The residual variance (after fitting) was 0.305 range units, which compares with 0.4 found with the variate difference method and with the 0.25 that was built into the data illustration. Note that reference (a) also covers the analysis of variance method of determining the residual variance after fitting. This is used as the estimate of the precision of the measuring system, assuming a smooth, low-order polynomial track.

c. Both the variate difference method and the orthogonal polynomial method are usable if a smooth track e.g., a low-order polynomial, is assumed and if the data are equally-spaced in the independent variable. If the data are not equally-spaced, then standard regression methods for low-order polynomial fitting can be used. Separate regression calculations must be made for each degree polynomial, testing each one for statistical significance. After the highest-order fit deemed significant statistically is determined, the residual analysis of variance, after fitting, is taken as the precision of the measuring system. Reference (a) covers polynomial fitting, statistical significance, and residual analysis of variance.

Reference (a): Natrella, Mary, Experimental Statistics. National Bureau of Standards Handbook 91, Oct 1966.

PAGE 6-10, REVERSE, BLANK

Section 7

Sampling From a Finite Population

701. Discussion

a. Most of the analysis effort usually concerns a relatively large population. However, at times the items to be tested become a large part (10 or 20%) of the entire population. In such cases the usual formulae, applicable to large populations, do not apply.

b. For illustration, an evaluation concerned an ejector seat for an aircraft trainer. The criterion was a success rate of at least 0.90. We decided to test to a 90% confidence level. Using a binomial confidence limit approach, we would have to test 22 ejector seats (without a failure). Only 16 ejector seats were to be used with actual aircraft between now and 1985. Thus the test sample would be a material part of the population. And the probability of drawing a defective unit would change as the sample is drawn.

c. The hypergeometric distribution should be used with the binomial.

d. In general, if a lot contains S units, m of which are defective, the probability that a random sample of N units will contain c defective units ($c \leq N$ or $c \leq m$ if $m \leq N$) is

$$P(\tfrac{c}{N}) = \frac{C_{N-c}^{S-m} C_c^m}{C_N^S}$$

e. Applying this to the illustration in a trial-and-error fashion we obtain, say for a test sample of 17:

$N = 17$
$S = 33$
$m = (.1)(33) = 3.3$ (Use as 3 defectives)
$c = 0$

$$P\left(\tfrac{0}{17}\right) = \frac{C_{17}^{30} C_0^3}{C_{17}^{33}} = .10$$

f. Thus, we are 90% confident if we obtain no failures in 17 trials. Thus, our test sample size should be 17 in lieu of 22 if we had ignored the finite population modification.

Section 8

Sequential Testing

801. Introduction

a. Sequential testing is a procedure that can materially reduce the needed number of runs or firings. Prior to actual testing, a chart is prepared based on certain criteria. It is divided into three decision areas. During testing the result of each firing is plotted and a decision is made.

(1) Continue with the test.

(2) Discontinue and accept.

(3) Discontinue and reject.

b. The purpose of the test runs must be resolved into a simple yes/no answer. Is the average detection range at condition A for the new model better than 14,000 yd, the average range of the old model? Is the hit percentage more than 45%? Is the CEP less than 55 ft? Another requirement is that the result of each run or firing be known before another run is made.

c. Sequential testing is useful mainly in the "technical" type testing that must often be done in our evaluations prior to our beginning the performance type determinations. The result is confidence in our yes/no answer. There is no result as to how much better or worse.

802. Illustration

a. Criteria

(1) We want to accept the equipment at the 0.90 probability level if the CEP is 2.3 kft or lower.

(2) We want to reject the equipment at the 0.90 probability level if the CEP is 3.5 kft or higher.

(3) If the CEP is between 2.3 and 3.5 kft is is immaterial whether we accept or reject, i.e., it is a special case.

b. Decision Rules

(1) Based on the above criteria the decision process is derived. (See attached plot.) The average expected sample size can also be calculated. This depends, of course, on what the CEP really is. If the CEP is actually zero, this process would lead to acceptance in five firings. If the CEP is really 2.3, the

sample size is expected to be 13 firings on the average. If the CEP is really 3.5, the average sample size is 8.

(2) Figure 8-1 gives the sequential plotting of an artificial series of firings. (See Table 8-1). Note that no decision has been reached yet in this illustration. If the miss distance to be observed on the 18th firing is less than 1.5, then this firing would lead to a decision. Definite stoppage rules can be determined to limit the total number of firings.

Table 8-1

Illustration of Sequential Testing

Number of Firing	Miss Distance (R)	$(R)^2$	Cumulative R^2
1	1.2	1.4	1.4
2	2.1	4.4	5.8
3	0.1	0.0	5.8
4	2.8	7.8	13.6
5	0.5	0.2	13.8
6	0.8	0.6	14.4
7	1.3	1.7	16.1
8	4.7	22.1	38.2
9	3.9	15.2	53.4
10	1.2	1.4	54.8
11	2.0	4.0	58.8
12	0.8	0.6	59.4
13	1.8	3.2	62.6
14	0.5	0.2	62.8
15	2.7	7.3	70.1
16	2.0	4.0	74.1
17	1.2	1.4	75.5

NOTE: Illustration made up from situation with CEP = 3.0

(3) The details of determining the limiting lines are given in WALD, A - "Sequential Analysis", New York, Wiley 1947. This reference also gives procedures for sequential testing using binomial data, quantitative data, etc.

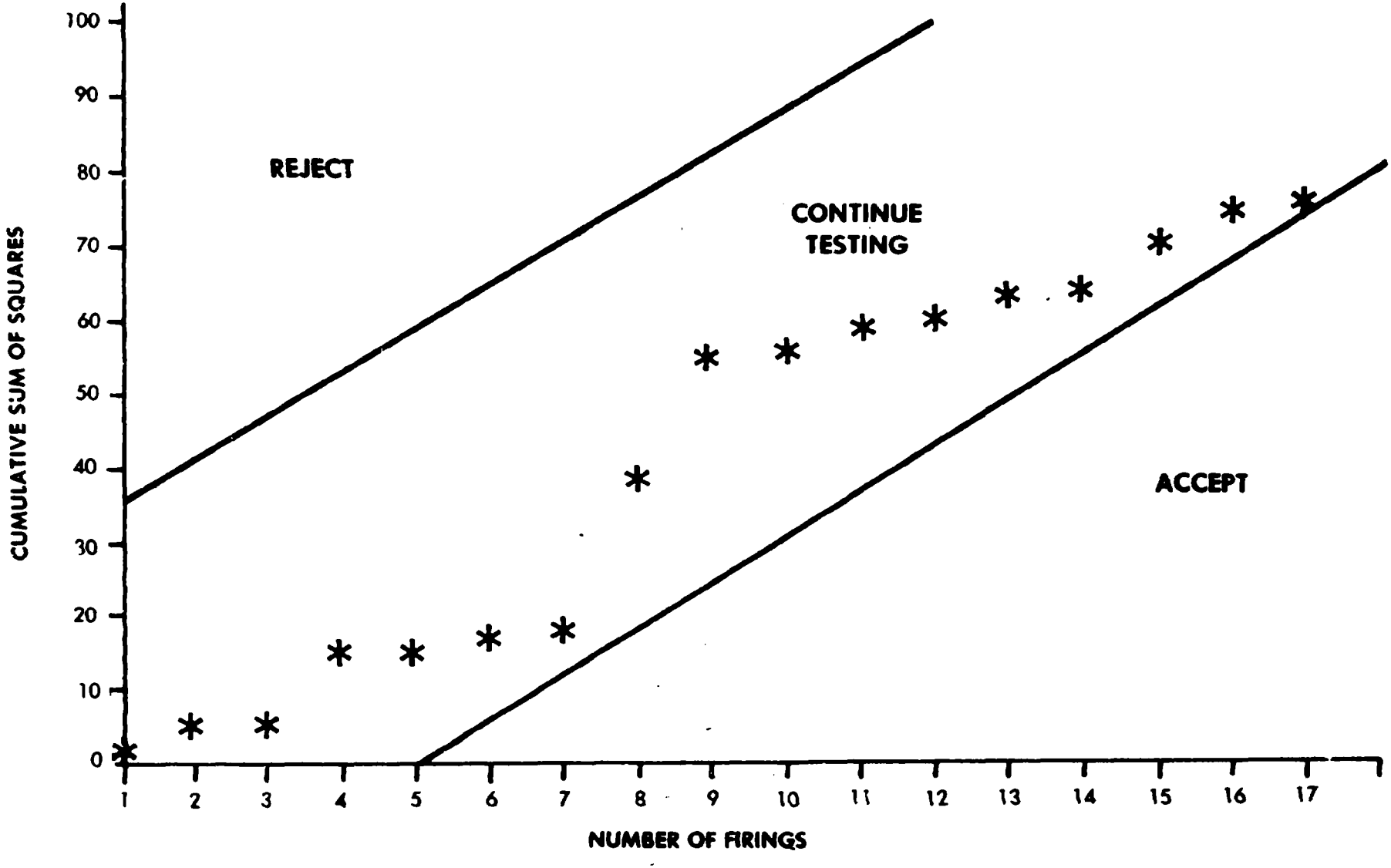


FIGURE 8-1
ILLUSTRATION OF SEQUENTIAL TESTING

Section 9

Use of Components of Variance

901. Discussion

a. A recent project involved aircraft tracking the target and obtaining a fire control solution for a weapon drop. Each aircraft used three types of displays. These were found similar to each other in bearing error. The analysis of variance technique was used to obtain the following measures of accuracy. See a statistical text on how to determine components of variance.

(1) Reading Error (s_r): This measured the variation in reading the three displays at the same time on an aircraft. The standard deviation in bearing was 6°.

(2) Tracking Error (s_t): This measured the variation in accuracy readings taken 5 minutes apart on an aircraft during the period of tracking. If the reading error was zero, the standard deviation in tracking was 2°. This represented the effect of the uncontrolled variables that changed over a short period of time.

(3) Sortie Error (s_c): This measured the sortie-to-sortie variation in accuracy, independent of the reading and tracking errors. It pertained to the ground checks and calibration procedures as they affected the accuracy. It was 5° (standard deviation).

b. To determine the accuracy of the fire control solution for a weapon drop, the three errors are combined as follows:

$$\sqrt{\frac{s_r^2}{r.t.c} + \frac{s_t^2}{t.c} + \frac{s_c^2}{c}}$$

Where r is the number of readings in each tracking period, t is the number of tracking periods, c is the number of aircraft simultaniously tracking the same target and pooling their results. With the intended tactics to be used with the system, c is always unity.

c. If the accuracy had to be improved, this tells which of the three errors should be worked on: sortie error (s_c). In the meantime we must live with the error values as determined. We should devise tactics that drop the weapon after one or two tracking time periods.

Section 10

Confidence Limits on MTBF

1001. **Discussion**

a. If we have assumed an exponential reliability function and a constant failure rate, we may use the Chi-square distribution to determine limits for a chosen confidence level. Table 10-1, which gives the single-sided, lower MTBF confidence limit factors, has been included to simplify the confidence limit calculations. The use of this table is best illustrated by means of an example. Let it be assumed that during an evaluation, 1,000 hours of operating time were accumulated. During this period, 10 mission aborting failures were experienced. The observed MTBF is 100 hours. To determine the single-sided, lower confidence limit at the 80% confidence level, enter Table 10-1 at 10 test failures and 80% confidence level and read the lower single-sided limit factor of 0.73. Multiplying this factor by the observed MTBF yields the single-sided lower limit, 73 hours. As a result of the evaluation, it can be reported that the MTBF experienced during the evaluation was 100 hours and, with 80% confidence, that the true MTBF is greater than 73 hours.

b. The MTBF and associated confidence limit can be used to determine probability of successful mission life. This is using the basic reliability function formula for R(t). If the equipment has a mission life of 30 hours and an MTBF of 73 hours at the single-sided lower 80% confidence limit, then the probability of performing its mission without a failure is:

$$R(t) = e^{-\frac{30}{73}} = 0.66$$

(at the 80% confidence level).

c. An important feature of the exponential distribution is that in obtaining the total operating time for subsequent MTBF calculations, all of the items under test are used. For example, in a sonobuoy reliability evaluation, there were 10 sonobuoys under test. Six of the 10 operated for 12 hours before they were pulled out for some non-failure reason. Three of the other four ran for 24 hours before the test was discontinued (not due to failure). The remaining one ran for 3 hours before it failed. The total operating time would be (6 x 12) + (3 x 24) + 3 = 147 hours with one failure. The MTBF would be 147 hours.

d. The above feature of the exponential distribution warrants OPTEVFOR special attention as it points out the importance of having more than one item for test if at all possible at evaluation time.

Table 10-1

MTBF Multiplication Factors for Single-Sided Lower Confidence Limit Exponential Distribution*

*Extracted from "MTBF Confidence Limits", T. A. Simonds Industrial Quality Control, Vol. XX, No. 6 Dec 1963.

Number of Failures	Confidence Level(%)								
	50	60	70	75	80	85	90	95	99
0**	1.44	1.10	.83	.71	.62	.51	.44	.34	.22
1	.60	.50	.41	.37	.33	.30	.26	.21	.15
2	.75	.64	.55	.51	.47	.42	.38	.32	.24
3	.82	.72	.63	.59	.54	.50	.45	.39	.30
4	.86	.76	.68	.64	.60	.55	.50	.44	.34
5	.88	.79	.71	.66	.62	.57	.52	.46	.36
6	.90	.82	.74	.70	.66	.62	.57	.51	.41
7	.92	.83	.76	.72	.68	.64	.60	.53	.44
8	.92	.85	.78	.74	.70	.66	.62	.56	.46
9	.93	.86	.79	.75	.72	.68	.63	.57	.48
10	.94	.87	.80	.77	.73	.69	.65	.59	.50
12	.95	.88	.82	.79	.76	.72	.67	.62	.52
14	.96	.90	.83	.80	.77	.73	.70	.64	.55
16	.96	.90	.85	.82	.79	.75	.71	.66	.57
18	.97	.91	.86	.83	.80	.76	.72	.68	.59
20	.97	.92	.87	.84	.81	.77	.74	.69	.60
25	.97	.92	.88	.86	.83	.80	.77	.72	.64
30	.98	.93	.89	.86	.84	.81	.78	.74	.66
50	.99	.95	.92	.90	.88	.85	.83	.79	.72
100	.99	.97	.94	.93	.92	.90	.88	.85	.79

** For zero failures since no MTBF can be determined, the factor is used with total test time; this permits a statement that "the MTBF is greater than" some value.

e. A more detailed description of MTBF confidence limits is given by T.A. Simonds (Industrial Quality Control, Vol. XX, No. 6)

f. The basic formulas are dependent upon the specific test plan utilized to collect the data upon which the limits are based: (a) fixed time truncated life test; (b) fixed failure truncated life test.

g. A fixed time truncated life test is one in which one or more equipments are put on test until a certain amount of total test time has elapsed (with failed parts and/or equipments being replaced when failures occur), this total test time being the summation of the times that each equipment on test has operated. Once this designated amount of total test time has elapsed, the test is stopped and the total number of test failures occurring up to then is recorded.

h. A fixed failure truncated life test is one in which one or more equipments are put on test until a certain total number of test failures have occurred, this total number of test failures being the summation of the individual numbers of test failures occurring on each equipment. Once this designated total number of test failures has occurred, the test is stopped and the total elapsed test time is recorded.

i. For our purposes, we have assumed that the fixed time truncated life test is the basis for establishment of the desired confidence limits since this test is utilized much more frequently than the fixed failure test. Even within this framework, there is some question as to the degrees of freedom for the Chi-square distribution. D. R. Cox uses $2r + 1$ as the degrees of freedom based on a heuristic study. The following is more standard:

(1) Double-Sided Limits:

$$\theta_U = 2T/\chi^2\ (1-\alpha/2,\ 2r)$$

$$\theta_L = 2T/\chi^2\ (\alpha/2,\ 2r + 2)$$

(2) Single-Sided Limits:

$$\theta_U = 2T/\chi^2\ (1-\alpha,\ 2r)$$

$$\theta_L = 2T/\chi^2\ (\alpha,\ 2r + 2)$$

(3) In these formulas:

θ_U = Upper MTBF Confidence Limit;

θ_L = Lower MTBF Confidence Limit;

T = Total test time.

r = Total number of test failures occurring during the test.

α = Risk level to be used in determining the MTBF confidence limit(s), which is equal to 1 - C.L., where C.L. is the desired confidence level. (NOTE: Caution should be exercised when looking up values for α in tables. Table 10-2 gives χ^2 in the form $P(\chi^2 > x) = \alpha$. For 80% confidence level, 8 degrees of freedom, the .200 column should be used, giving 11.03.

χ^2 = Chi-square value associated with the subscript quantities listed in parentheses for the four formulas. The first quantity; listed variously as: $1 - (\alpha/2)$, $\alpha/2$, $1 - \alpha$, and α; is the proportion of the chi-square distribution lying above the desired value for χ^2. The second quantity; listed variously as: $2r + 2$, and $2r$; is the number of degrees of freedom to be applied in the determination of a value for χ^2. Once these two quantities are known, the appropriate value for χ^2 can be selected. (NOTE: If a fixed failure truncated life test were to be used, the only change would be that the number of degrees of freedom associated with the χ^2 values used in determining the lower confidence limits would be 2r instead of 2r + 2.)

TABLE 10-2 CHI SQUARE TABLE

$$P(\chi^2 \geq x) = \alpha$$

ν	.995	.990	.980	.975	.950	.900	.800	.750	.500	.250	.200	.100	.050	.025	.020	.010	.005
2	.01	.02	.04	.05	.10	.21	.45	.58	1.39	2.77	3.22	4.60	5.99	7.38	7.82	9.22	10.59
4	.21	.30	.43	.48	.71	1.06	1.65	1.92	3.36	5.39	5.99	7.78	9.49	11.15	11.66	13.28	14.82
6	.67	.87	1.13	1.24	1.63	2.20	3.07	3.45	5.35	7.84	8.56	10.65	12.60	14.46	15.03	16.81	18.55
8	1.34	1.64	2.03	2.18	2.73	3.49	4.59	5.07	7.34	10.22	11.03	13.36	15.51	17.55	18.17	20.08	21.94
10	2.15	2.55	3.06	3.24	3.94	4.86	6.18	6.74	9.34	12.55	13.44	15.99	18.31	20.50	21.17	23.19	25.15
12	3.06	3.57	4.18	4.40	5.22	6.30	7.81	8.44	11.34	14.85	15.81	18.55	21.03	23.35	24.06	26.25	28.25
14	4.07	4.65	5.36	5.62	6.57	7.79	9.47	10.16	13.34	17.12	18.15	21.07	23.69	26.13	26.88	29.17	31.30
16	5.14	5.81	6.61	6.90	7.96	9.31	11.15	11.91	15.34	19.3	20.47	23.55	26.30	28.86	29.64	32.03	34.32
18	6.25	7.00	7.90	8.23	9.39	10.86	12.86	13.68	17.34	21.61	22.76	25.99	28.88	31.54	32.35	34.83	37.21
20	7.42	8.25	9.23	9.59	10.85	12.44	14.58	15.45	19.34	23.83	25.04	28.42	31.42	34.18	35.03	37.59	40.05
22	8.62	9.53	10.59	10.98	12.34	14.04	16.31	17.24	21.34	26.04	27.30	30.82	33.93	36.79	37.67	40.31	42.84
24	9.87	10.85	11.99	12.40	13.84	15.66	18.06	19.04	23.34	28.24	29.56	33.20	36.42	39.38	40.28	43.00	45.60
25	11.13	12.19	13.40	13.84	15.38	17.29	19.82	20.84	25.34	30.44	31.80	35.57	38.89	41.94	42.86	45.66	48.33
26	12.44	13.55	14.84	15.30	16.92	18.94	21.59	22.66	27.34	32.62	34.03	37.92	41.34	44.47	45.43	48.30	51.04
30	13.77	14.94	16.30	16.79	18.49	20.60	23.36	24.48	29.34	34.80	36.25	40.26	43.78	46.99	47.97	50.91	53.71
32	15.10	16.35	17.78	18.29	20.07	22.27	25.15	26.30	31.34	36.97	38.47	42.59	46.20	49.50	50.49	53.51	56.37
34	16.48	17.78	19.27	19.80	21.66	23.95	26.94	28.13	33.34	39.14	40.68	44.91	48.61	51.98	53.00	56.08	59.00
36	17.86	19.21	20.78	21.33	23.26	25.64	28.73	29.97	35.34	41.30	42.88	47.22	51.00	54.45	55.50	58.64	61.62
38	19.26	20.68	22.29	22.87	24.88	27.34	30.54	31.81	37.34	43.46	45.08	49.53	53.39	56.91	57.98	61.18	64.22
40	20.67	22.14	23.82	24.42	26.51	29.06	32.35	33.67	39.34	45.61	47.26	51.80	55.75	59.34	60.44	63.71	66.80
42	22.10	23.63	25.37	25.99	28.14	30.77	34.16	35.52	41.34	47.76	49.45	54.08	58.12	61.78	62.90	66.23	69.37
44	23.55	25.12	26.93	27.56	29.79	32.49	35.98	37.37	43.34	49.91	51.63	56.36	60.48	64.20	65.34	68.73	71.93
46	25.01	26.63	28.49	29.15	31.44	34.22	37.80	39.23	45.34	52.05	53.81	58.63	62.83	66.62	67.78	71.22	74.47
48	26.48	28.15	30.07	30.75	33.10	35.95	39.63	41.09	47.34	54.19	55.99	60.90	65.17	69.03	70.20	73.70	77.00
50	27.96	29.68	31.65	32.35	34.76	37.69	41.46	42.95	49.34	56.33	58.16	63.16	67.50	71.42	72.62	76.17	79.52
52	29.45	31.22	33.24	33.96	36.44	39.44	43.29	44.81	51.34	58.46	60.32	65.41	69.83	73.81	75.03	78.63	82.03
54	30.95	32.77	34.85	35.58	38.12	41.19	45.12	46.68	53.34	60.59	62.49	67.67	72.15	76.20	77.43	81.09	84.53
56	32.46	34.33	36.45	37.20	39.80	42.94	46.96	48.55	55.34	62.72	64.65	69.91	74.46	78.57	79.82	83.53	87.03
58	33.98	35.89	38.07	38.84	41.49	44.70	48.80	50.42	57.34	64.85	66.81	72.15	76.77	80.94	82.21	85.97	89.51
60	35.50	37.46	39.69	40.47	43.19	46.46	50.65	52.30	59.34	66.98	68.97	74.39	79.08	83.30	84.59	88.40	91.98
62	37.04	39.04	41.32	42.12	44.89	48.23	52.49	54.18	61.34	69.10	71.12	76.62	81.38	85.66	86.96	90.82	94.45
64	38.58	40.63	42.95	43.77	46.59	50.00	54.34	56.06	63.34	71.22	73.27	78.85	83.67	88.01	89.33	93.23	96.91
66	40.13	42.22	44.59	45.42	48.30	51.77	56.19	57.94	65.34	73.34	75.42	81.08	85.96	90.35	91.69	95.64	99.36
68	41.68	43.82	46.23	47.08	50.02	53.55	58.05	59.82	67.34	75.46	77.56	83.30	88.25	92.69	94.04	98.04	101.80
70	43.25	45.42	47.88	48.75	51.74	55.33	59.90	61.70	69.34	77.57	79.71	85.52	90.53	95.03	96.39	100.44	104.24
72	44.81	47.03	49.54	50.42	53.46	57.12	61.76	63.59	71.34	79.69	81.85	87.74	92.81	97.36	98.74	102.83	106.68
74	46.39	48.65	51.20	52.10	55.19	58.90	63.62	65.48	73.34	81.80	83.99	89.95	95.08	99.68	101.08	105.22	109.10
76	47.97	50.27	52.86	53.78	56.92	60.69	65.48	67.37	75.34	83.91	86.13	92.16	97.35	102.00	103.42	107.60	111.52
78	49.55	51.89	54.53	55.46	58.65	62.49	67.35	69.26	77.34	86.02	88.27	94.37	99.61	104.32	105.75	109.97	113.94
80	51.14	53.52	56.20	57.15	60.39	64.28	69.21	71.15	79.34	88.13	90.40	96.57	101.88	106.63	108.07	112.34	116.35
82	52.74	55.16	57.88	58.84	62.13	66.08	71.08	73.04	81.34	90.23	92.53	98.77	104.14	108.94	110.40	114.71	118.75
84	54.34	56.80	59.56	60.53	63.88	67.88	72.95	74.94	83.34	92.34	94.66	100.97	106.39	111.25	112.72	117.07	121.15
86	55.95	58.44	61.25	62.23	65.62	69.68	74.82	76.83	85.34	94.44	96.79	103.17	108.65	113.55	115.03	119.43	123.55
88	57.56	60.09	62.93	63.91	67.37	71.49	76.69	78.73	87.34	96.54	98.92	105.37	110.90	115.84	117.35	121.78	125.94
90	59.17	61.74	64.63	65.64	69.13	73.29	78.56	80.63	89.33	98.65	101.05	107.56	113.14	118.14	119.65	124.13	128.32
92	60.79	63.39	66.32	67.35	70.88	75.10	80.44	82.53	91.33	100.75	103.17	109.75	115.39	120.43	121.96	126.48	130.71
94	62.41	65.05	68.02	69.06	72.64	76.92	82.31	84.43	93.33	102.85	105.30	111.94	117.63	122.72	124.26	128.82	133.08
96	64.04	66.71	69.72	70.78	74.40	78.73	84.19	86.33	95.33	104.94	107.42	114.13	119.87	125.00	126.56	131.15	135.46
98	65.67	68.38	71.43	72.50	76.16	80.54	86.07	88.23	97.33	107.04	109.54	116.31	122.11	127.28	128.85	133.49	137.83
100	67.30	70.05	73.14	74.22	77.91	82.36	87.95	90.14	99.33	109.14	111.66	118.49	124.34	129.56	131.15	135.82	140.19

PAGE 10-6, REVERSE, BLANK

Section 11

Sample Size, Comparison of Two Means

This section gives a method to determine needed sample size to compare two test conditions. Normal distribution is assumed. Sections 28, 29, and 30 also present sample size methodology for the more general case of meeting a threshold value.

1101. The Formula Approach (Simple Comparison)

a. There are formulae to relate sample size (N), adequacy (risk factors α and β), estimated experimental error (s), and magnitude in results deemed operationally important (Δ). Specifying any four of the five elements, the fifth one can be determined.

(1) What difference in results are we interested in? If you consider that difference (Δ) between, say, modes are operationally important if results differ by 10 miles, our sample size needs are much smaller than if we wish to detect differences of 3 miles.

(2) What is our expected experimental error? If our standard deviation (s) is expected to be small, our sample size need not be so large.

(3) What risk are we willing to tolerate in incorrectly deciding that we have an operationally important difference? If our toleration level is high ($\alpha = 0.15$), our needs are lower than if we demand more confidence ($\alpha = 0.01$).

(4) What risk are we willing to tolerate in missing an operationally important difference? If our toleration level is high ($\beta = 0.10$), our needs are lower than if we are more stringent ($\beta = 0.05$). Or if our confidence (P) in detecting a significant result must be high, our sample size must be large. See Table 11-1.

Table 11-1

Relationship Between Sample Size and β Errors

Number of Replications	Probability of β Error
2	.92
3	.83
4	72
5	62
7	41
10	20
15	05
20	02

Based on a constant α error ($\alpha = .05$ and $s/\Delta = 1$).

1102. Illustration I

a. The formula procedure concerns five parameters (N, α, β, Δ, s). We can solve for N by fixing the other four. However, these formulae will be illustrated with the usual situation in our work, that is, with the amount of services, N, fixed. We need then determine the adequacy of our sample size.

b. Suppose we have a simple evaluation to compare mode 1 and mode 2. Our available sample size is 16 runs each. The standard deviation from a previous evaluation was 3.0 miles. (This was based on 25 degrees of freedom.) Now let's quickly run through the procedure to get a feeling for what's involved.

(1) Step 1: Calculate standard deviation of difference in means.

$$s_{\bar{X}_1 - \bar{X}_2} = s\sqrt{\frac{1}{\bar{N}_1} + \frac{1}{\bar{N}_2}} = 3.0\sqrt{\frac{1}{16} + \frac{1}{16}} = 1.06 \text{ miles}$$

(2) Step 2: We enter the t-table to find the t value for 25 degrees of freedom and the α probability level. We have selected α as being equal to 0.05. Then, the t value is 2.06, which is

$$2.06 = \frac{\bar{X}_1 - \bar{X}_2}{1.06}$$

$$\text{or } \bar{X}_1 - \bar{X}_2 = (1.06)\ (2.06) = 2.18 \text{ miles}$$

So our observed difference must be 2.18 for detection as a significant difference at the $\alpha = 0.05$ level.

(3) Step 3: The probability of actually obtaining a difference of 2.18 or larger depends on the true difference of the two modes. Form the t-statistic

$$t = \frac{\Delta - 2.18}{\bar{X}_1 - \bar{X}_2}$$ where Δ is the true difference.

(4) Step 4: If Δ is set equal to 5, then

$$t = \frac{5 - 2.18}{1.06} = 2.8$$

(5) Step 5: We enter the t-table with 25 degrees of freedom and our t-value of 2.8. The corresponding probability level (β) is found to be 0.01 with $1-\beta=P=0.99$.

(6) Step 6: Conclusion: We are confident (P=0.99) of detecting a true difference of 5 miles as being significant in our setup with sample sizes of 16 each.

c. The procedure can be repeated for various values of the different elements; or graphs can be used to aid in the final selection with respect to trade-offs of the various parameters.

1103. Illustration II

a. When our critical data are quantitative and normal, the relationship among the risk factors is shown as follows:

Sample Size Factor

Using significance level, (α), in analysis	Probability of obtaining a significant result ($1-\beta$)		
	0.80	0.90	0.95
0.20	10.5	14.8	18.8
0.10	13.9	18.8	23.7
0.05	17.5	23.4	28.1

b. To determine the number of replications (r) needed per mean, say, in comparing two modes, we multiply the appropriate cell value by $(s/\Delta)^2$. For example, if:

(1) A difference of 50 units or larger is deemed operationally important, and

(2) We expect the standard deviation to be 40 units, and

(3) We will use a significance level of 0.05 in analysis, and

(4) We want to have a 0.90 confidence in detecting a significant result, then

(5) The number of data going into the mean of each of the two modes must be at least $(40/50)^2$ (23.4) = 15.

c. Table 11-2 is a corresponding sample size table for binomial type data. Only a few situations are given: $1-\beta$ of .80 and 0.70 and Δ of 20% for one mode and 70, 80, and 90% for the contrasting test mode.

d. There are many sample size formulae depending on criteria, situation and assumptions. A useful reference in this regard is Mary Natrella's Experimental Statistics, National Bureau of Standards Handbook 91, Oct. 1966. Sample size formulae, tables, etc. are given for many situations. This is an excellent reference book for many data analysis techniques and experimental designs. Another handy tool for the analyst is a reliability sliderule, called a Reliability Computer. This, among other things, determines sample size needs for a binomial and also for MTBF. OPTEVFOR analysts can obtain one from 02B in Norfolk (8-690-5177).

e. The sample size formula in this section pertains to comparison of two samples (two means, etc.). For sample size needs for single sample means, MTBF, etc., see Sections 28, 29, and 30.

1104. The Total Sample Concept. The formula approach is a somewhat naive approach to sample size determination. For example, the factorial approach in our projects does not involve sample sizes of 8 or 16 per condition. Actually, the sample size is influenced by every step in the project plan, there are many considerations that have more influence on the sample size than those appearing explicitly in the above formulae.

a. Are your data quantitative or qualitative? The cost and effort involved in taking measurements of a continuous type, say miss distances, pays off in reducing our sample size needs over a simple count of hits/misses. If we had to use only hit/miss type data, the sample size needs would have to be enlarged considerably.

b. Is the analytic approach a piecemeal, one-at-a-time approach, or is it more involved, integrating many conditions in a balanced fashion (factorial type)? The sample size per condition would be less if we had a larger-scope integrated evaluation than if we were content with just a single piece of information. This is the total sample concept. The concept leads to: with a large total sample we can test as many variables as necessary, running per condition only 2,1, or even 1/2 sample size.

(1) The sample size factor table and the sample size formula procedure are still pertinent. However, they pertain not to

Table 11-2

Sample Size Per Mode for Count Type Data

The values below illustrate the sample sizes needed with hit/miss type data. Read as follows: If we wish to be 80% confident in detecting a significant difference when one Mode is only 20% and the other is 70%, we would need 15 trials per mode if we use a significance level of .05.

If we wish to be 80% confident in detecting a significant difference when one Mode is only 20% and

	the other Mode is		
	70%	80%	90%
and the analyst uses a significance level of			
.05	15	10	7
.10	12	8	6
.15	11	7	5
.20	9	6	5

If we wish to be 70% confident in detecting a significant difference when one Mode is only 20% and

	the other Mode is		
	70%	80%	90%
and the analyst uses a significance level of			
.05	12	8	6
.10	9	7	5
.15	8	6	4
.20	7	5	4

Note: Two tail

the per condition, but to the highest order interaction table expected to be significant in the evaluation. For example, suppose two variables interacting was the highest level expected and this would be a table of, say, four means, each based on eight data points. For our purposes we can then consider the "significance" of the evaluation to be based on a sample size of eight.

(2) A measure of experimental error will be available by pooling high order interactions in the analysis. For example, most analysts pool all high-order interactions into a remainder or experimental error term.

(3) Pooling high order interactions into an error term usually gives a more meaningful and useful error measure than that obtained from repeat runs. For example, duplicate or triplicate readings are concerned with measurement error, which is of little consequence compared to the other causes of variability in our evaluations.

(4) Though not testing each condition completely, results at each condition can still be obtained. This is based on the statistical model used and determined in the analysis.

(5) Basically the question of the analytic approach boils down to: is it better to test a few conditions well or to use the limited services to test many conditions? In our operational evaluations, using the total sample concept, we generally test many conditions to cover the broad scope needed.

Section 12

Sensitivity Testing (Up and Down)

1201. Discussion

a. The up-and-down method is useful when the test data are either successful or non-successful depending on the stress level of the variable under test. For example, while the amount of explosive is a continuous variable, the lethal amount or threshold cannot be measured directly, say with a single explosion. Usually the lethal dose (LD_{50}) is found by testing a prescribed number of explosions at each of several fixed dose levels. In the up-and-down method, the dose levels tested are determined sequentially. A sample size saving usually results. The steps in the up-and-down method are:

(1) A series of test levels is chosen with equal spacing between doses (equal spacing on the appropriate scale, usually log-dose). This spacing is chosen approximately equal to the guesstimated standard deviation.

(2) A series of trials is carried out following the rule of an increase in dose following a response (O) and a decrease in dose following a non-response (X). The first test should be performed at a level as near as possible to the LD_{50}.

(3) The number N' is the total number of tests performed in each series. Testing continues until a chosen "nominal" sample size N is reached. N is the total number of trials reduced by one less than the number of like responses at the beginning of the series. This accounts for the poor first test level. For a series OOOXXOXO, we have N'=8 and N=6.

(4) The resulting configuration of responses and non-responses for each series is referred to the table of maximum likelihood solutions (Table 12-1) for the LD_{50} and one computes

$$X_f + kd$$

where X_f is the last dose administered, k is the tabular value and d is the interval between doses. Table 12-1 lists all solutions for all N' and $N \leq 6$. If the series begins with X vice O, the sign of k is to be reversed.

(5) See Figure 12-1 for an example of a test series.

Table 12-1

VALUES OF k FOR ESTIMATED LD_{50}

N	Second Part of Series	k for Test Series Whose First Part is 0	k for Test Series Whose First Part is 00	Standard Error of LD_{50}
2	X	-.50	-.39	.88σ
3	X0	.84	.89	.70σ
	XX	-.18	.00	
4	X00	.29	.31	.67σ
	X0X	-.50	-.44	
	XX0	1.00	1.12	
		.19	.45	
5	X000	-.16	-.15	.61σ
	X00X	-.88	-.86	
	X0X0	.70	.74	
	X0XX	.08	.17	
	XX00	.30	.37	
	XX0X	-.30	-.17	
	XXX0	1.29	1.50	
	XXXX	.56	.90	
6	X0000	-.55	-.55	.56σ
	X000X	-1.25	-1.25	
	X00X0	.37	.38	
	X00XX	-.17	-.14	
	X0X00	.02	.04	
	X0X0X	-.50	-.46	
	X0XX0	1.17	1.24	
	X0XXX	.61	.73	
	XX000	-.30	-.27	
	XX00X	-.83	-.76	
	XX0X0	.83	.94	
	XX0XX	.30	.46	
	XXX00	.50	.65	
	XXX0X	-.04	.19	
	XXXX0	1.60	1.92	
	XXXXX	.89	1.33	

Source: W.J. Dixon, JASA, Dec 1965

Section 13

CPD (Cumulative Probability of Detection)

1301. Introduction. A typical measure of detection, say in radial closing runs, is to define opportunity of detection and then form the ratio of detections to opportunities. This can be put in cumulative form, say by range. The following is a more convenient way to obtain CPD that eliminates the troublesome definition of opportunity. In doing so, this method uses all the information in the runs. This method is the conditional probability of detection method.

$$CPD_i = CPD_{i-1} + P_i\ (1\text{-}CPD_{i-1})$$

It is conditional in that the cumulative probability of detection by a certain range or range bin i is equal to the cumulation prior to the i range or i bin plus the probability of being detected in the i bin times the number entering the i bin undetected. Note that a detection made by a particular range is assumed to likewise be detections at all shorter ranges.

1302. Example

a. Five radial closing runs went as follows:

Run	Comex	Detection
1	4000	3275
2	3000	2000
3	5000	No detection (CPA = 2600)
4	3500	2704
5	5000	2500

b. The first step is to obtain P_i which takes the number detected to available into account.

Range Value	P_i	$1\text{-}CPD_{i-1}$	CPD_i
3275	1/4	0	.25
2704	1/4	.75	.44
2600	0/3	.56	.44
2500	1/2	.56	.72
2000	1/1	.28	1.00

c. Note that this procedure can handle different Comex ranges. This is seen in the P_i value for 3275 range of detection which is 1/4 since one Comex did not occur by 3275 range. The assumptions for this procedure are that range to the target is the key variable for detection. Continuously exposed for detection is also assumed. Note also that this procedure leads to a 1.00 value whenever the shortest range event is a detection.

PAGE 13-2, REVERSE, BLANK

Section 14

Analysis of Unbalanced Data

1401. General

a. Out-of-Line Data

(1) The first step in analysis is to determine out-of-line data. For example, if all our data points are between 3 and 8 kyd except one which is 18 kyd, the latter is definitely suspect. The environment and actual test conditions would be examined. Possible technical causes should be ferreted out if possible. If no satisfactory explanation is forthcoming, the problem becomes an analysis problem: detection of out-of-line data and then how to handle them. While there are formulae and procedures available to determine which data are out-of-line, these are not too helpful in our work. Here is an approach used on a recent project.

(2) In analyzing bearing accuracy, the data set contained some "wild" readings. These were considered valid, however. Some preliminary analysis indicated a standard deviation of 15°. A decision was made to divide the data into two types: readings greater than ±60° and readings less than ±60°. The latter set was analyzed using standard techniques. The former were considered "failures" for analysis purposes, and an overall "failure" rate (4%) was found. The analysis on both sets were combined in the results reported. Thus, precision measures were reported. Note that this procedure does not throw out any data. It merely divides the data for different analytical procedures in order to facilitate the analysis.

b. Combining With Count Data

(1) A procedure similar to the preceding is useful in combining, say, range at detection data with runs having no contact. Nature has divided the runs into two types. The ranges at detection, when available, are analyzed using the normal or log-normal distribution. The no-contact rate is studied to determine if the rate varies by condition. The results, say a CPD curve by range, are presented combining both analyses.

(2) If, in the preceding, the reason for the no-contact runs are considered due to run geometry prematurely truncating the run (CPA), then "truncated" or "censored" statistical procedures can be used.

c. Unbalance

(1) Our most prevalent difficulty in analysis is unbalance caused by missing data. In analyzing ranges at detection we

may have many no-contact runs. In analyzing miss distances we may have many direct hits. Runs will be aborted, data recorders will malfunction, services will be chopped, etc.

(2) Obviously in some projects the occurrence of the missing data, say the no-contact runs, may be the most important result in the report. The analysis problem still remains: how do we analyze the unbalanced set of data, say the detection ranges.

(3) If test conditions were selected using the piecemeal or threat approach, balance was not deemed important and the simple analysis procedures could still be used as planned. With a factorial type approach balance is very important. Missing data usually seriously complicate the analysis.

(4) If just a few data points are missing (less than 5%), missing plot formulae can be used to fill in the blank cells, and the analysis procedures can be used in a straightforward manner. If the missing data are mainly restricted to one level of a variable, say, altitude less than 200 feet, the scope of the analysis using standard techniques can be restricted.

(5) Most likely, many data points will be missing more or less randomly throughout the matrix of test conditions. In these cases a multivariate regression (paragraph 504) technique should be used. Basically the statistical model is explicitly formed with perhaps 30 or so terms including many interactions. Special coding is used to handle the non-quantitative settings. Computer programs are available to determine the significance or non-significance of the various terms in the model. The best fit to the data is obtained and is used to predict the results at the various test conditions. Our experience with this technique is limited, and its uses and limitations are still being studied. For example, we think that this procedure will also be useful when the data are qualitative (count of hits/misses or of detections, non-detections). Its use in such cases should be supplementary only.

d. Non-Normal

(1) If the data are definitely non-normal, all of the analysis techniques based on normal are weakened considerably. The significance probability value, say, may really be lower than calculated with our analysis technique in such cases. Or, perhaps, the converse.

(2) Transformations may be useful to transform a non-normal distribution to a normal one. The log transformation has already been discussed (paragraph 302). The arcsine or logit has been used with precentage data. The square root transformation for count type data has also been used.

(3) In some cases the situation may need more data processing, for example, analysis of radial miss errors in bombing drops. Radial errors, disregarding relative angle, are certainly non-normal. And no transformation is useable. The procedure used is not to analyze the radial values directly but to continue the data processing by determining the X-Y component for each radial. The analysis is then done on the X component separate from the Y component. Results are combined for final presentation. The separate results are also important, particularily when the coordinates are relative to the flight path.

(4) Non-parametic analysis techniques are available for non-normal situations. These are practically distribution-free. The median (the middle data in rank) is often used in lieu of the mean. Associated tests of significance are available with the median. Ranking tests such as the sign rank test or the sign test are also available for making simple test of significance. Mary Natrella's Experimental Statistics, National Bureau of Standards Handbook 91, Oct 1966 has good coverage.

1402. Simple Case (See also paragraph 402)

a. In evaluating system acquisition time against seven threats, 16 runs were available in "good" environments and 16 in "poor" environments.

Threats						
1	2	3	4	5	6	7
Good Environment						
81	42	60	55	63	71	94
74	53	57	40	66	77	84
				64		105
Poor Environment						
77	45	61	50	82	73	101
94	80	83	58		87	
	38	65	42		96	

b. The initial step in data analysis is to standardize the data for the different environments. One way to proceed is to average all 16 good environment runs (67.9) and average all poor environment runs (70.8). However, this is wrong in this case because of the unbalance of runs by threat and environment. For example, those averages would be strongly influenced by three runs for threat 7 in good environment and one run in poor environment. A better way is to find the average time for each threat and environment set. The environment effect can be found by finding the differences as below.

Averages

Good Environment

77.5	47.5	58.5	47.5	64.3	74.0	94.3

Poor Environment

85.5	54.3	69.7	50.0	82.0	85.3	101.0

Environment Effect (Good minus Poor)

-8.0	-6.8	-11.2	-2.5	-17.7	-11.3	-6.7

An assumption can be made that the environment effect is the same for all seven threats. The average difference is -9.2 sec. (Note that this is quite different from 67.9 - 70.8 = -2.9 if the unbalance was ignored.) The -9.2 environment effect can be used to standardize the data for, say, good environment. By adding -9.2 seconds to each poor environment data point, we remove environment as a variable in the analysis. The analysis can then proceed by averaging each threat, etc.

Threats (In good Environment)

1	2	3	4	5	6	7
81	42	60	55	63	71	94
74	53	57	40	66	77	84
68*	36*	52*	41*	64	64*	105
85*	71*	74*	49*	73*	78*	92*
29*	56*	33*		87*		

* Adjusted

1403. Complicated Case. Step-wise regression (paragraph 504) and normal equation formulae and solutions are standard ways to analyze a complicated set of unbalanced data. These procedures involve computer usage with many assumptions.

Section 15

Confidence in an Observed Proportion

1501. Discussion. When the data are counts (e.g., the number of hits), NBS Handbook 91* gives graphs and tables useful in determining confidence limits about the observed percentages. Tables A-22 and A-23 give tables of one- and two-sided confidence limits for samples of 30 or less. Table A-24 is a series of charts useful for confidence limits when sample sizes are large. For example:

a. A sample of 100 rockets were tested.

n=100. n is the size of the sample.

b. 60 of the rockets operated satisfactorily.

r=60. r is the number of operable rounds.

r/n=0.6 is the satisfactory fraction observed from the sample.

c. At the 95% confidence level the acceptable fraction of the population may be expected to lie within the range 0.50 to 0.70 as shown on page T-46*.

d. To use the curve, read the values of p where the curves for n=100 intersect the vertical for r/n=0.6. The values are 0.50 and 0.70. This means that it is 95% certain that from 50% to 70% of the rocket population will be good, or that some value between 30% and 50% of the rocket population will be duds. A more precise way of stating this is: If a large number of samples of n=100 were taken and each sample were similarly tested, 95% of the confidence intervals computed would contain the true value.

* Experimental Statistics by Mary Natrella, Oct. 1966. National Bureau of Standards Handbook 91.

Section 16

Using Confidence Intervals as Tests of Significance

1601. Discussion. Prof. D. Barr, Naval Postgraduate School, relates confidence intervals to tests of significance between two means (Journal of Quality Technology, Oct 1969).

a. When separate 100 $(1-\alpha)$ percent confidence intervals on two population means do not overlap, the means are significantly different at the 100α percent level. The assumptions are similar to those for the z test (large sample sizes). When the intervals do overlap, the means may or may not be significantly different at the 100α percent level. Prof. Barr gives a "two-interval" test to determine the significance. Consider the null hypotheses H_0: $\mu_1\mu_2=0$ against the two-sided alternative. The test statistic is

$$\text{Test } z = \frac{\bar{X}_1-\bar{X}_2}{\alpha} \sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

and the test criterion is to reject H_0 when $z<z_{\alpha/2}$ or $z>z_{1-\alpha/2}$. This is formally equivalent to constructing a $100(1-\alpha)$ percent confidence interval for $(\mu_1 - \mu_2)$. To be equivalent to the usual test for equality of means, the intervals for the two-interval test must have confidence coefficient γ less than $(1-\alpha)$. The exact confidence coefficient is $\gamma = 2F(q) - 1$ where $F(q)$ is the cumulative normal distribution of γ, where

$$q = \frac{\sqrt{(n_1 + n_2)}}{\sqrt{n_1} + \sqrt{n_2}} \quad (z_{1-\alpha/2})$$

b. Table 16-1 exhibits this relationship for a number of values of α.

Table 16-1

	Significance Test	Two-Interval Test
α	$z_{1-\alpha/2}$	q
0.20	1.283	0.639
0.10	1.645	0.758
0.05	1.960	0.837
0.01	2.575	0.933

c. For example, if two 83.7% confidence intervals don't overlap, the difference between the two means is signifi cant at the 5% level.

d. See also Chapter 21 in the National Bureau of Standards Handbook #91.

Section 17
Combining Probabilities from Independent Tests of Significance

1701. Discussion

a. Consider the case when independent tests of significance do not lead to significance individually. If all observed comparisons are in the same direction, the question is the significance of the aggregate. If the data are the same for the various tests, some combined method, such as ANOVA, could be used. If data are of different kinds, then we must use the individual probabilities or significance levels.

> "The circumstance that the sum of a number of values χ^2 is itself distributed in the χ^2 distribution with the appropriate number of degrees of freedom, may be made the basis of such a test. For in the particular case when n = 2, the natural logarithm of the probability is equal to $-1/2\chi^2$. If therefore, we take the natural logarithm of a probability, change its sign and double it, we have the equivalent value of χ^2 for 2 degrees of freedom. Any number of such values may be added, together, to give a composite test, using the Table of χ^2 to examine the significance of the result."

The above is a direct quote from R. A. Fisher, Statistical Methods for Research Workers, 1949.

b. Suppose command and control Method B was observed to be better than A in an ASW barrier situation. The improvement in terms of targets engageable was at the significance level of 0.145. In an AA situation, Method B had more aircraft engaged, significance level of 0.087. Method B also had a slight edge in number of missiles engageable, significance level of 0.263. In testing the aggregate (unweighted in this case) we have:

α	$-\log_e \alpha$	Degrees of Freedom
.145	1.93	2
.263	1.33	2
.087	2.44	2
	5.70	6

$$\chi^2 = 2 (-\ln \alpha) = \chi^2 = 11.42$$

The observed χ^2 value of 11.42 is compared with the tabular value for six degrees of freedom. The 5% value is 12.592; the 10% value is 10.645. The probability of the aggregate of the three tests occurring by chance therefore is not far from 0.075.

c. See also Littell and Folks "On the Comparison of Two Methods of Combining Independent Tests", JASA, Mar 1972

Section 18

Determining P_k for Salvos from Single Firings

1801. Discussion

a. To reduce the number of missile firings, OPTEVFOR often determines P_{k1} (kill probability for single firings) and then calculates the corresponding P_{k2} for salvos of two. If independence could be assumed, then the classical formula applies.

$$P_{k2} = 1 - (1-P_{k1})^2$$

b. A recent project involved a system that fired two missiles at the same target within a few seconds. Both firings used the identical fire control solution. The above formula would not apply: The proper method includes separation of the P_{k1} single firing test data into

P_{fcs} "successful" FCS

P_m "successful" missile flight

then

$$P_{k2} = P_{fcs} \{1 - (1-P_m)^2\}$$

The situation is likely to be more complex than the modified formula indicates. But the extensions should be straightforward.

Section 19

Calculation of Multi-Events

1901. Discussion

a. CDR Thurneysen used a very efficient compact procedure to calculate multi-event probabilities. The procedure is use of a matrix; but first, the situation which led to this illustration:

(1) Scenario: Four helos are available to transport 16 loads from ship to shore. Transport time is 0.6 hr ship-to-shore and 0.6 hr shore-to-ship. One helo can make as many as six trips during the time available. What is the probability of completing this scenario (from the reliability standpoint). The mean time between mission-aborting failures is 18.9 hr ship-to-shore. Not all the engines, etc., need be used from shore to ship. The corresponding mean time between mission-aborting failures is estimated to be 20.0 hr shore-to-ship.

(2) The reliability: ship-to-shore is 0.969 and shore-to-ship is 0.970. The probability of a helo making a complete trip is 0.940. The corresponding figures give the results for one helo. (See Table 19-1)

b. Table 19-2 is a work table using the inputs directly from Table 19-1 as column and row headings. The cell entries are the product correspondingly of these headings. The diagram below Table 192 illustrates how to use the table. The diagonals give the cells to sum to obtain the probability of making 0, 1, 2, . . . 12 trips with two helos. Example: to determine the probability of making 10 trips, use diagonal labeled 10. This diagonal includes cell 6, 4, cell 5, 5, cell 4, 6. The entries .0334 + .0019 + .0334 sum to .0687. This is the probability of making 10 trips with two helos in our scenario.

c. The sum of each diagonal is now posted as column and row headings in Table 19-3. Again diagonals are formed. Since the scenario is keyed to 16 trips, only the diagonals 16 - 24 trips need be used. Similar to Table 19-2, cell entries are formed. The sum of all entries noted in Table 19-3 is 0.8853. This is the probability of completing the scenario.

Table 19-1

ONE HELO SUMMARY

Number of Trips	Formula	Value
0	q	.060
1	pq	.056
2	$p_2 q$	.053
3	$p_3 q$	.050
4	$p_4 q$	.047
5	$p_5 q$	.044
6	$p^6 q$	.711*

* Does not include shore-to ship on last trip.

TABLE 19-2

TWO HELO SUMMARY

NUMBER OF TRIPS: ONE HELO

Number of trips: In Helo		0 .060	1 .056	2 .053	3 .050	4 .047	5 .044	6 .711
0	.060	0036	0034	0032	0030	0028	0026	0313
1	.056	0034	0031	0030	0028	0026	0025	0398
2	.053	0032	0030	0028	0027	0025	0023	0377
3	.050	0030	0028	0027	0025	0024	0022	0356
4	.047	0028	0026	0025	0024	0022	0021	0334
5	.044	0026	0025	0023	0022	0021	0019	0313
6	.711	0313	0398	0377	0356	0334	0313	5055

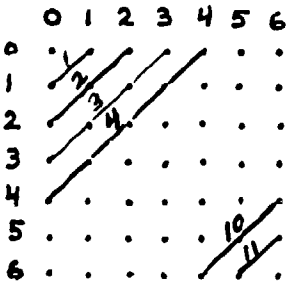


TABLE 19-3

FOUR HELO SUMMARY

Number of trips: Two Helos

		0	1	2	3	4	5	6	7	8	9	10	11	12
		0036	0068	0095	0120	0140	0158	0751	0890	0820	0754	0687	0626	5055
0	0036													
1	0068													
2	0095													
3	0120													
4	0140													0071
5	0158												0010	0080
6	0751											0052	0047	0380
7	0890										0067	0061	0056	0450
8	0820									0067	0062	0056	0051	0415
9	0754								0067	0062	0057	0052	0047	0381
10	0687							0052	0061	0056	0052	0047	0043	0347
11	0626						0010	0047	0056	0051	0047	0043	0039	0316
12	5055					0071	0080	0380	0450	0415	0381	0347	0316	2555

Section 20

Some Goofs in Data Analysis

2001. Introduction. This chapter gives some examples of poor analysis. While each example is based on a real situation, the sample sizes, mean values, etc. have been changed drastically to make an impact and to stress the point.

2002. Confounding Because of Unbalance

a. When the sample sizes vary among the different test levels (unbalanced), the observed means of each test variable alone may be misleading. Table 20-1 and 20-2 illustrate a situation.

Table 20-1

Table of Means by Pulse Variable

Pulse	Sum	N	Mean
With	2250	25	90
Without	1380	23	60

Table 20-2

Table of Means by Device

Device	Sum	N	Mean
With	2550	35	73
Without	1080	13	83

b. If the above averages were significantly different, one may conclude that effectiveness increased by 30 units with pulsing and decreased by 10 with the Device. However, as the two variable table of means, Table 20-3, indicate, this result would be wrong.

Table 20-3

Table of Means by Pulse and Device

	Device								
Pulse	With			Without			Both		
	Sum	N	Mean	Sum	N	Mean	Sum	N	Mean
With	1350	15	90	900	10	90	2250	25	90
Without	1200	20	60	180	3	60	1380	23	60
Both	2550	35	73	1080	13	83			

c. As Table 20-3 indicates, the use or non-use of the Device has no effect on the mean values. With Pulse the means (90) are the same with and without the Device. Without Pulse the means (60) are the same with and without the Device. Thus, the apparent difference in the overall Device means, 73 versus 83, are merely due to the unbalance in sample sizes indicated in Table 20-3.

d. What to do to handle the above type of situation? If unbalance cannot be avoided deal with at least two-variable tables. If the project concerns many variables, use more formal techniques of analysis such as step-wise regression.

2003. Separating Performance and Reliability

a. Obtaining the MOMS directly without initial separation of performance and reliability is poor analysis practice. Table 20-4 gives results that indicate an apparent superiority of Tactics B over Tactics A.

Table 20-4

MOMS Worktable by Tactic

Description	Tactic A	Tactic B
	Number of Runs	
Hardware Failures	40	10
Targets Killed	40	50
Non Killed	20	40
Total	100	100
MOMS	.40	.50

The MOMS calculation (# targets killed/total) gives 0.40 for Tactic A and 0.50 for Tactic B.

b. Rather than a direct calculation, let's first consider the situation. In this illustration, as in most of our work, the hardware failures are not a function of the particular variable or condition under test. For example the hardware (or software) may not "know" that a low-level attack is scheduled. Or the stress on the hardware is the same whether it's low- or high-altitude attack. Thus, combining all hardware failures gives a better failure rate estimate than each tactic separately.

c. Based on Table 20-4 the failure rate is (40+10)/ (100+100) or 0.25. The hardware success rate is 1-.25 or .75. This is the probability of the system being up. The performance, given that the hardware (and software) is up, is measured by # targets killed/# killed and non-killed. For Tactic A it is 40/60; for Tactic B it is 50/90.

d. Table 20-5 gives the MOMS Worktable. Notice that the MOMS is quite different from that reported in Table 20-4.

Table 20-5

MOMS Worktable by Tactic

	Tactic	
Measure	A	B
Performance When Up	.67	.57
Hardware Being Up	.75	.75
MOMS	.50	.42

2004. Combining Horses and Cows. A typical blunder is fostered by thinking that "this is exactly what we observed at sea in our testing. We made 200 classifications and 100 were correct. Our classification accuracy is 50%. Since this is what we got, it must be right." Suppose the figures are as in Table 20-6.

Table 20-6

Classification Accuracy Worktable

Type	# Correct	# Total	%
Diesel	75	100	75
Nuke	25	100	25
Both	100	200	50

Note that the overall 50% figure is not pertinent to diesel type or to nuke type. Since the accuracy is significantly different by type, an overall figure has meaning only if the weights are pertinent to the expected weights in real combat. What is not pertinent are the weights based on sample size. If we expect in, say, 1982, the diesel/nu ratio to be 20% to 80%, then the overall is (.75)(.20) + (.25)(.80) = .35 or 35%.

Section 21

The Bayesian Approach

2101. Introduction

a. In T&E a great deal of information is obtained by others. And, of course, we use the results of this early testing, not only in our early involvement but also in our OPEVAL planning and testing. This is usually done in an informal manner, using classical analytical techniques. This paper presents in elementary form the Bayesian Approach that formally combines previous test results and results of later testing into a final result. A savings in amount of testing usually occurs when the Bayesian Approach is used.

b. There is strong disagreement among analysts as to the proper use of the Bayesian Approach. This is based on axiomatic differences in part. Regardless, the Bayesian Approach should be considered at OPTEVFOR as a formal analysis tool. It may be especially useful "under the table", particularly in OT-III. This is one reason for describing the approach. The other is that contractors, etc. do use it. We should know the basics of the approach when we deal with them.

2102. Simple Illustration. The example concerns an automatic sensing device; the critical data are the percentage of false alarms. (Adapted from reference a.)

a. The first step is to express formally our prior information. This is called our prior or a priori distribution. In this simple example only three false alarm rates are considered possible or likely, centered at 0.10, 0.15, and 0.20. The OTD guesstimates that the proportion of these ratios occurring are corresponding 0.60, 0.25, and 0.15. (Our best guess averages 0.13 at this stage.) See Table 21-1, Columns 1 and 2.

b. The next step is to make tests at sea. The OTD was able to make 20 runs during which he got 4 false alarms. (Note: 4/20 = 0.20.) The Likelihood Column in Table 21-1 gives the odds in getting exactly 4 out of 20 if the "true" percentage was 0.10, 0.15, and 0.20, respectively.

c. The next column, Joint Probability, in Table 21-1, is also a working column, a product of the previous two columns.

d. The last column in Table 21-1 is the ratio of each Joint Probability entry to the sum. The Posterior Probability is a revision of our prior estimates based on the at-sea results. This Posterior Probability is the Bayesian result. Our best estimate at this stage averages 0.14. The revision in the aver-

age from 0.13 to 0.14 (insignificant in this example) is due to the added information (4/20 = 0.20) from the runs at sea. Our final answer with this approach (0.14) is quite different from using just the runs at sea (0.20). The larger the disparity between our a priori estimates and the results at sea, the larger the influence of the results at sea. The larger the test at sea, the larger the influence of the at-sea results.

e. If our prior distribution is non-informative or non-discriminatory, then the final result is based on the at-sea results. For example, if the prior in this illustration included all possible false alarm rates between 0 and 1, each having the same probability of occurring, then the final result is based on the test at sea. This indicates the nature of the classical or non-Bayesian Approach. Standard or classical statistical methods presupposes all possible values as possible and likely.

Table 21-1

Simple Illustration

Prior Distribution			Joint	Posterior
Event	Probability	Likelihood	Probability	Probability
p_i	$P_o(p_i)$	$P(x=4/p_i)$	$P_o(p_i)P(X=4/p_i)$	$P(p_i(X=4))$
.10	.60	.0898	.0539	.41
.15	.25	.1821	.0455	.34
.20	.15	.2182	.0327	.25
	1.00		.1321	1.00

Note: $P(x=4/p_i) = \frac{n!}{x!(n-x)!} p_i^{x} (1-p_i)^{n-x}$

Where x is number of false alarms.

$P(p_i/x=4) = P_o(p_i)P(X=4/p_i)/P(x=4)$

2103. A Materiel Reliability Example. This example concerns the materiel reliability (MTBF being the critical measure) of a complex electronic equipment such as a sonar system. No real data are available since the program is in OT-I stage.

a. The first step in the Bayesian Approach, formation of our prior distribution, in this example is based entirely on judgment as flavored by the Incentive Fee Contract. Figure 21-1 gives the OTD guesstimates as to the odds of getting such MTBF values. (Note that specific odds are not necessary; the relative odds in graphic form are sufficient.) The specific odds are then found

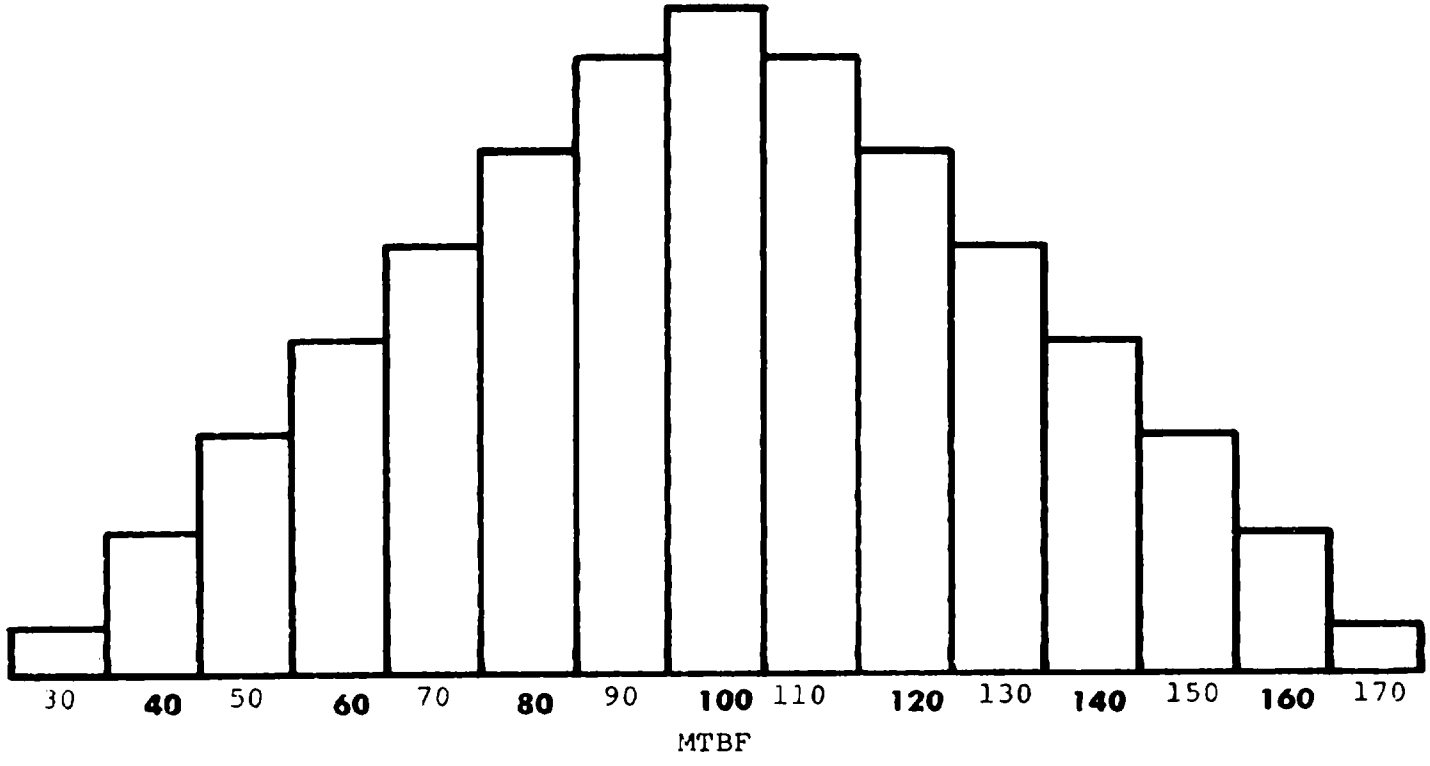


FIGURE 21-1

MTBF PRIOR DISTRIBUTION

by proportioning the number of cells per MTBF zone to the total number of cells under the curve. Table 21-2 gives the result of this proportioning (Column 2) by MTBF midpoint (Column 1).

b. At this stage, a test at sea or ashore reveals 14 failures in 523.3 hours, for a non-Bayesian MTBF of 37 hours. Note that this would have been published in our report as the result if we were using the usual non-Bayesian Approach. With the Bayesian Approach, this is merely more information to use to revise our a priori guesstimates.

c. The Likelihood column in Table 21-2 gives the probability of obtaining 14 failures in 523 hours if the actual event (MTBF) was correspondingly 30, 40 ... 170 hours. The first row value is interpreted as: the probability is 0.0737 that we would observe 14 failures in 523 hours given that the true MTBF is 30 hours. (The values are found by analytical formulae, exponential in this case.)

d. The Joint Probability column in Table 21-2 is the product of the previous two columns. As such, then, it is the probability, say the first value of 0.0007, of obtaining 14 failures in 523 hours and the true MTBF is 30 hours.

e. The Posterior Probability is merely the corresponding values of Joint Probability compared to the sum of the Joint Probability values. The initial value of 0.07, say, is the probability of the true MTBF being 30 hours given that 14 failures were observed in 523 hours.

f. The final result, the Posterior Column, gives 53 hours as the MTBF. (The Prior Distribution gave 100 hours as the MTBF.)

2104. Discussion

a. In addition to formal use of prior information, the Bayesian Approach differs from the classical in some concepts. For example the Bayesian analyst considers probability as a degree of belief in a statement. A classical approach considers a true, fixed value of a population mean and derives confidence intervals. These intervals are random with known probability of containing the true value. A Bayesian talks about a probability interval. This is the interval in which lies the parameter value that is the random variate.

b. The Bayesian Approach is keyed to the selection of the prior distribution.

(1) In the example, the Bayesian MTBF of 53 hours is higher than the at-sea result of 37 hours because the prior distribution averaged higher than the at-sea average. If the prior distribution had averaged lower than the at-sea result,

Table 21-2

MTBF Illustration

Prior Distribution				Joint	Posterior
Event	Cell Count	Prob.	Likelihood	Probability	Probability
30	50	.0089	.0737	.0007	.07
40	150	.0268	.1027	.0028	.29
50	250	.0446	.0617	.0028	.29
60	350	.0625	.0275	.0017	.18
70	450	.0804	.0110	.0009	.09
80	550	.0982	.0043	.0004	.04
90	650	.1161	.0017	.0002	.02
100	700	.1250	.0007	.0001	.01
110	650	.1161	.0003	.0000	.00
120	550	.0982	.0001	.0000	.00
130	450	.0804	.0001	.0000	.00
140	350	.0625	.0000	.0000	.00
150	250	.0446	.0000	.0000	.00
160	150	.0268	.0000	.0000	.00
170	50	.0089		.0000	
	5600	1.0000		.0095	

Note: The Likelihood, based on the exponential, is

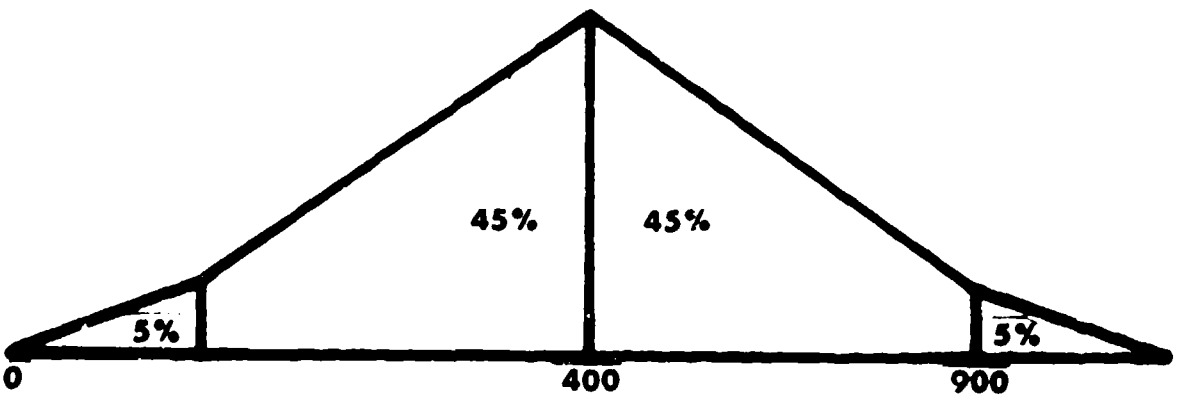
$$P(x/\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$ where λ is failure rate and x is number of failures.

then the Bayesian result would have been lower than the at-sea result.

(2) As an extreme, it is possible to select the prior distribution in such a way that no at-sea test would be warranted.

(3) The selection of the prior distribution is subjective. Even is the data were available, such as DT-I data, the impact of operational conditions vice technical would be subjective.

(4) Formal mathematical distributions such as Gamma or Beta are used extensively in the literature as priors. In a mathematical sense these priors work very well. However, for our needs a graphical prior is sufficient and is not too difficult to obtain from the OTD. For example, the OTD may be willing to guesstimate three values for a prior: real low result (5%), real high result (95%) and the center result (50%). A workable prior could be formed, as:



(5) While the prior parameters are importa.t, the shape of the distributions are not to the final result. Figure 21-2 is from reference b. Fig 21-2a, (n=0) illustrates two different priors. N is normal, u = 1, s = 1. E is exponential with parameter 1 also. Figures 21-2(b), (c), (d), (e) are resultant posteriors with test runs samples

n = 1, n = 4, n 9, n = 25 respectively.

The test runs were all sampled with $\bar{x}$ = 1, s = 1. Under these conditions differences in prior, even with small samples (n=4), have little effect on end results.

2105. Savings. The Bayesian Approach usually saves services. An evaluation can usually be conducted with less firings, etc using Bayesian vice classical; Dr. Bonis' (Ref c) gives examples. Say for a binomial case considering the number of trials with 0 failures: to demonstrate a reliability of 0.90, the classical

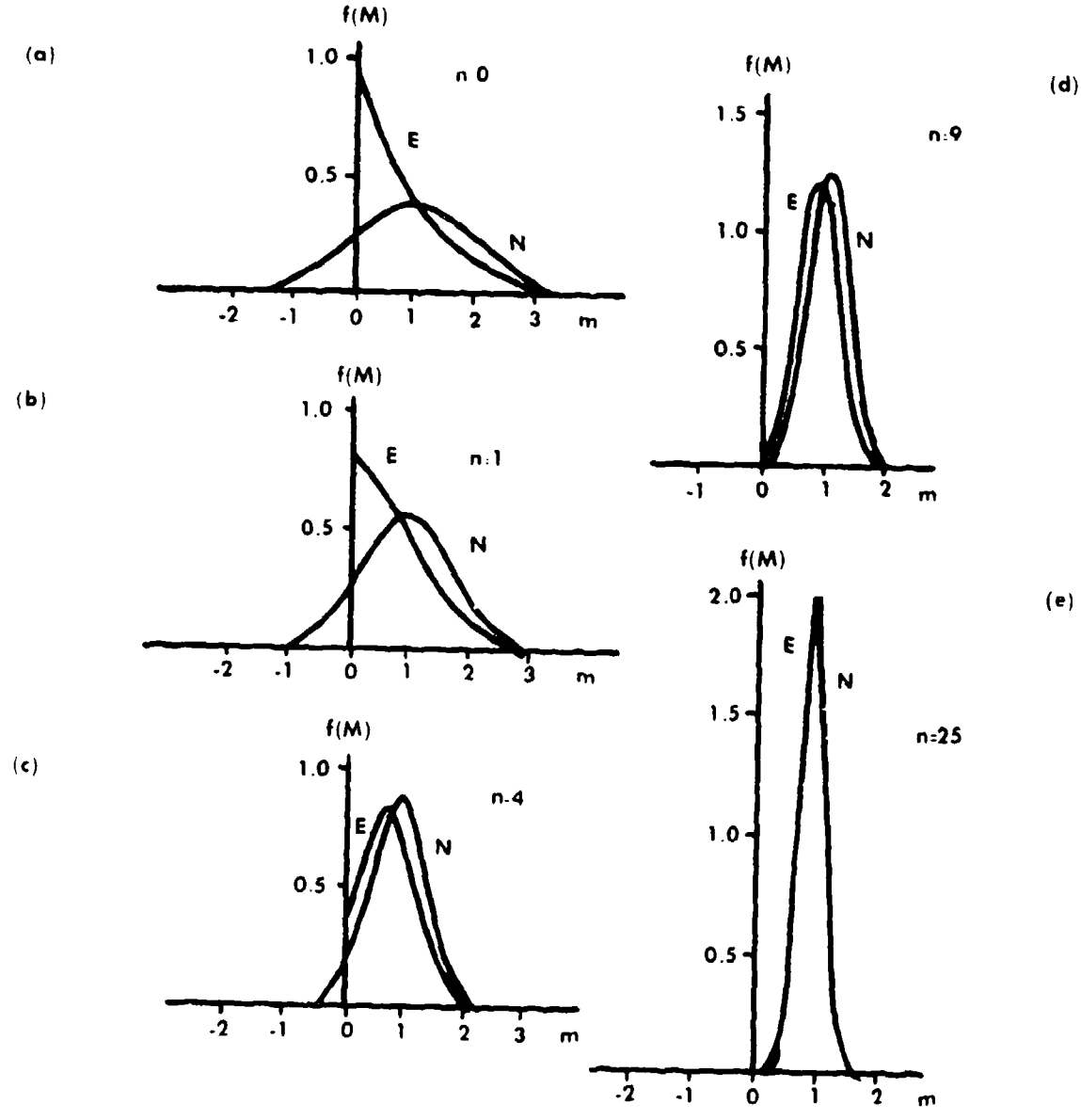


Figure 21-2

Posterior Distributions Corresponding to Normal and Exponential Prior Distributions

approach calls on 22 to 44 firings, depending on the confidence. Using the Bayesian, the sample size is 3 to 43 depending on the confidence and prior. Dr. Bonis shows with a very simple prior we need 8 firings without a failure to demonstrate that the reliability is at least 0.80 at the 90% confidence. The classical approach needs 11 firings. The prior that Dr Bonis used is simply that the reliability of 0.80 has a 50% chance of being met. See Figure 21-3.

2106. More Discussion

a. There is no doubt that as evaluators we need to minimize the subjective element. There is no doubt that our use of the Bayesian Approach, i.e., selection of the prior distribution, should be limited. When prior information is available, say TECHEVAL data, then we may decide to use this approach.

b. We may be forced to use this approach in certain situations, due to testing limitations. Even if highly subjective in nature, it may be worthwhile.

c. Sensitivity analysis should be done on each project before use. For example, how much variation in the prior can we tolerate before final results are affected? Should we use worse case and best case and combine in some pert-like fashion? How much saving in testing does the approach actually give us?

2107. References

a. Morgan, Bruce W., An Introduction to Bayesian Statistical Decision Processes. Prentice-Hall, Inc., 1968.

b. Britt, P. S. and Ibbotson, E. L., A Bayesian Approach to Determining the Sample Size For Maintainability Demonstration. Thesis, Air Force Institute of Technology, June 1969.

c. Bonis, A. J., Why Bayes is Better. Proc, 1975 Annual Reliability and Maintainability Symposium.

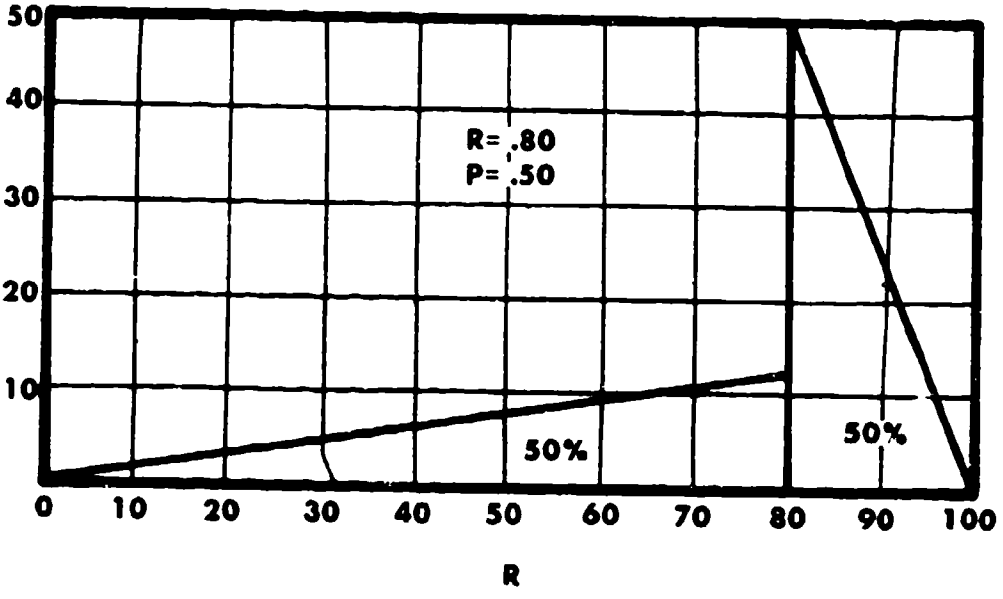


Figure 21-3

PAGE 21-10, REVERSE, BLANK

Section 22

Reliability When Failure Times Are Not Known Exactly

2201. Introduction

a. An important operational characteristic is captive-carry availability of missiles and torpedoes. Captive-carry availability is the reliability of weapons when carried on an aircraft, ship, or submarine for long periods of time. Captive-carry life is a form of shelf-life (sometimes under stress and/or at some level of energizing).

b. A captive-carry program has unusual features affecting analysis.

(1) The times to failure are not known exactly but only within inspection intervals.

(2) Even after a failure, unfailed components of the system are continued to be stressed until the end of the inspection interval.

(3) The sample may be censored, i.e., some unfailed missiles may be removed from the testing program for firing purposes, etc.

(4) The failed items may or may not be replaced.

c. An exponential distribution is assumed. Various methods had been proposed to obtain the basic analysis parameter, MTBF. Discussions were held but no agreement had been reached. A Monte Carlo simulation was used to analyze the different methods of calculating MTBF. This paper reports the insight gained during that Monte Carlo simulation.

d. Each method is illustrated with the following captive-carry situation and data. The captive-carry test began with 30 missiles with inspection periods every 50 hours. If a component failed, the missile would be considered failed. However, this could only be determined at each inspection period. If a component failed, the remaining components in the missile were still stressed until the end of the respective 50, 100, 150, 200, 250, or 300 hr inspection period. Failed missiles were not replaced. After each inspection period, if two unfailed missiles were available, they were removed from captive-carry and fired. These were not replaced. The captive-carry terminated at 300 hours.

Time to Failure Data (hours of carry):

4	36	85	166
18	43	90	201(207)
19	47	110	231
25	60(61)	158	293

The first number above gives the actual failure time of the missile due to the initial component failure. Additional sets of digits in parentheses give corresponding failure of the second, third, etc., missile components. While the above data give the actual time of failure, for Monte Carlo purposes, the failure of a component and missile was known only at each inspection period, 50, 100, 150, 200, 250, 300 hours. Two missiles passing the individual inspection periods were withdrawn from testing by being fired. The test was stopped at 300 hours with one unfailed missile.

2202. Method A

Numerator: a) Total time on unfailed missiles plus

b) Time on failed missiles assuming failure occurred at midpoint of the respective inspection period

Denominator: Number of failed missiles

For sample data: MTBF = (2400 + 1625)/17 = 236.8 hours.

With I as the time between inspection periods, use of the midpoint (I/2) seems to be a good approximation for failure time (for the exponential). In Schneeiweiss's paper* the midpoint approximation is shown to be biased (leads to higher MTBF) by about I^2/ (12 MTBF). This was reflected closely in the Monte Carlo results. The Monte Carlo gave confidence in the use of midpoint except when the length of the period approached the expected MTBF. Method A is an operationally feasible method and has been used in our evaluations. However, it admittedly does not use all the information that is operationally available with respect to number of component failures. In this sense it is wasteful in that the sample size must be increased for the same amount of confidence. In addition to the bias caused by midpoint approximation, there is also the possibility of bias caused by ignoring additional component failures. There is some indication that there is a bias, though slight.

2203. Method B. This is the maximum likelihood method using the reiterative process based on CNA Memorandum 1105-75 by W.J. Heintzelman of 12 Aug. 1975.

For sample data: MTBF = 235.9 hours

*W. G. Schneeiweiss, IEEE Trans Rel. Vol R-25 #5 Dec 76.

This is an operationally feasible method. It is actually in use now at OPTEVFOR in some projects. It is similar to Method A in not using all available data on component failures, but is preferred over Method A when the length of the inspection period (I) is about the same order of magnitude as the expected MTBF. Otherwise Method A is preferred because of simplicity.

2204. Method C

Numerator: a) Total time on unfailed missiles

b) Time to end of the appropiate inspection period on failed missiles.

Denominator: Number of failed components

For sample data: MTBF = (2400 + 2050)/19 = 234.2 hours

This method, a strong contender for general use at OPTEVFOR, uses all information that is operationally available. The added information is that all of the unfailed components are still stressed even though some have failed. There is some bias in that the time to the next inspection period for the failed components is counted as operating time. The amount of this bias is small. The bias can be handled by reducing the numerator by $(I/2)\ \Sigma c/C$ where Σc is the total number of components determined to have failed. C is the total number of components (guesstimate) in the missile. The principal drawback to this approach is the guesstimate of the total number of components in a complex weapon system so that the likely bias in the method can be estimated.

2205. Method D

Numerator: a) Total time on unfailed missiles plus

b) Time on failed missiles assuming failure occurred at end of respective inspection period

Denominator: Number of failed missiles

For sample data: MTBF = (2400 + 2050)/17 = 261.8 hours

This is an operationally feasible method and was proposed by an outside agency. It is directly related to Method A except for an obvious bias of I/2 in each estimate of failure time. The results for this method was exactly 25 hours more than for Method A. This method need not be considered further.

2206. Conclusions

a. While the various methods do not differ widely, there is no general best method of determining reliability. The optimum method depends on the situation. In any case the likely amount of bias and loss can be estimated as a basis for selection.

b. If, after initial component failure, the remaining missile components are not under stress in the testing program, then use of Method A or B is recommended. Use Method A if the expected MTBF is at least somewhat larger than the inspection interval. Use Method B otherwise, when $I^2/(12\ MTBF)$ is large.

c. If, after initial component failure, the remaining components are still under stress in the testing program (until the missile is withdrawn), then use Method C. Calculate, if possible, the possible bias for insight.

Note: See also Nelson, W., "Optimum Demonstration Tests with Grouped Inspection Data from an Exponential Distribution" IEEE Trans on Reliability, Vol R-26 #3 Aug 1977.

Section 23

Reliability: Binomial or Exponential?

2301. Discussion

a. When the mission time (t) is known and failure rate is constant, there is a choice as to how the operational reliability is to be determined. The choice is:

(1) Binomial. Make a series of runs, each of length mission time (t) and the reliability is the ratio of successes (still operating at time (t)) to total runs. The analysis assumptions are fairly straightforward: ordinary Bernoulli trial, reliability is constant.

(2) Exponential. Determine the MTBF from total operating time over number of failures. Use the exponential to determine reliability (R) for mission time (t). The critical assumption is the exponential distribution or constant failure rate.

b. While the exponential has somewhat more stringent analysis assumption, it is easier to use in determining reliability in that the project operations need not be conducted for the exact mission times. This flexibility also pertains to determining reliability for different mission times.

c. It is well known that results are similar if each binomial trial is tested to mission time (t). Reliability determined by the binomial or by the exponential would be similar. Here is one proof deriving the binomial from the exponential.

$$R = e^{\frac{-t}{MTBF}} = 1 - \frac{t}{MTBF} + \frac{t^2}{MTBF}\ \frac{1}{2!} - \frac{t^3}{MTBF}\ \frac{1}{3!} + \ldots$$

In the above series expansion only the first two terms are important.

$$\text{So } R = 1 - \frac{t}{MTBF} = \frac{MTBF-t}{MTBF} = \frac{\frac{T}{f} - t}{\frac{T}{f}} = \frac{T-tf}{T}$$

where T = total test time
f = number of failures
s = number of successes
$\frac{T}{t}$ = number of mission time units = s+f

So $R = \frac{st + ft - ft}{st + ft} = \frac{s}{s + f}$, which is binomial reliability.

Section 24

Fractional Factorial Test Designs

2401. General. In COMOPTEVFOR's Project Analysis Guide (COMOPTEVFORINST 3960.8), the rudiments of experimental design are mentioned.

a. Split-Plot. In paragraph 209 the split-plot design is given as side-by-side or back-to-back. This is a useful operational test design and is the most popular for shootouts between competitors and for sensitive comparisons between two scenarios or tactics. By analyzing the data in terms of differences, environmental effects are automatically eliminated from the comparison. Take heed, however, in planning to use the side-by-side that the design depends on having both systems involved up at the same time.

b. Factorial. In paragraph 210, the power of the factorial is given. By integration of various test scenarios into one factorial, 16 threats or scenarios can be evaluated, say, for the same amount of testing needed for only seven threats without the factorial. There is no question that the factorial is extremely powerful. Each trial furnishes information on each scenario. This is the power of the integration. However, this feature is also its drawback. Because every piece of data is used in each result, a missed data point affects each result. Thus, control at sea to obtain valid data for each factorial condition is a must. If more than 10% of the data are missing, the power of the factorial is wiped out.

2402. Fractional Test Design (Including Squares)

a. The fractional design offers high promise of being our most powerful design. A fractional is simply say 1/2 or 1/4 or 1/9 of a complete factorial. A 2^6 factorial in a particular evaluation may lead to 64 different test conditions. However, say only $2^6/4$ or 1/4 of the conditions need be tested. In lieu of 64 tests, the same amount of useful information can be obtained from only 16 tests.

b. The fractional concept is illustrated by the following. To keep it simple, no experimental error is assumed in the data. A 2^3 factorial led to the following data.

	Pilot 1		Pilot 2	
	Slow	Fast	Slow	Fast
Medium	13	33	18	38
High	23	43	28	48

Analysis of the data indicates an overall average of 30.5 plus:

(1) There is an increase of 10 units for High runs over Medium runs.

(2) Pilot 2 has 5 more units than Pilot 1.

(3) Going fast leads to an increase of 20 units.

(4) There are no interactions.

c. Contrast the above with the following: Since there was no interaction between pilot and speed, pilot and altitude, and also between speed and altitude, only 1/2 of the factorial will be tested.

	Pilot 1		Pilot 2	
	Slow	Fast	Slow	Fast
Medium	13			38
High		43	28	

Analysis of the above four data points indicates an overall average of 122/4 = 30.5 plus:

(1) Medium averages 51/2 = 25.5; high averages 71/2 = 35.5, which gives an increase of 10 units.

(2) Pilot 2 averages 66/2 = 33 while Pilot 1 averages 56/2 = 28. Pilot 2 has 5 more units than Pilot 1.

(3) Slow averages 41/2 = 20.5; fast averages 81/2 = 40.5. Fast leads to a 20 unit increase.

d. Note that the answers obtained with four data runs are exactly the same as with eight runs. This, of course, is due to no interaction.

2403. Alias

a. Let's put the preceding fractional in another form.

	Slow	Fast
Medium	1	2
High	2	1

Note: This is easily recognized as a Latin Square, which is the most fraction of the fractionals.

b. In this form we see that the interaction of Speed (S) and Altitude (A) is determined by contrasting one diagonal with the other. That is, based on the data

	Slow	Fast
Medium	13	38
High	28	43

43 + 13 = 56 is contrasted with 28 + 38 = 66 for an average difference of 10/2 = 5. This is a measure of the interaction of Speed with Altitude. However, this is the same contrast and the same data treatment and the same answer for the Pilot effect. In other words, in the above fractional we have purposely confounded the main effect, Pilots, with an interaction, Speed with Altitude. We do not know nor can we determine how much of the resulting data contrast is due to Pilot or how much to the Speed with Altitude interaction. In the language of test design, the S with A interaction is "alias" with the P effect.

$$SA = P \text{ or the identity } I$$

$$I = SAP$$

Before we fractionalized, we had just a 2^2 or Speed with Altitude factorial with effects.

S
A
SA

After we fractionalize, we have a $2^3/2$. The effects are found by multiplying the 2^2 set of effects by the identity (I) using the algebra that all squared terms are equal to 1. Thus, the effects are:

$$I = SAP$$

$$S = S^2AP = AP$$

$$A = SA^2P = SP$$

$$SA = S^2A^2P = P$$

The above indicates that S is alias with AP, A is alias with SP and SA is alias with P. This is the price we pay for running a fractional in lieu of a complete factorial.

c. Obviously if we felt that the AP interaction did not exist, then S would be determined pure and direct. Similarly for the other effects. Or to put this another way, the fractional would not be useable or valid unless the interactions were not existent. These assumptions of no-interactions can be relaxed to relatively free of interactions depending on trade-offs, etc.

d. National Bureau of Standards Applied Mathematics Series 48 (15 April 1957), 54 (1 May 1959) and 58 (1 Sept 1961) give a cookbook approach to fractionals. They pertain to plans at two settings, three settings, and mixed respectively. The selection, the test conditions, the analysis, the limitations can be aided significantly with these cookbooks.

Section 25

Estimating Total Number of Software Errors

2501. Discussion

a. Glenford J. Myers in his book Software Reliability (John Wiley & Sons, 1976, page 338) gives an interesting method to estimate the number of errors in a software program. We need two groups independently testing the same set of test cases. After testing for a time period, the errors found by each group are listed. N_1 and N_2 are the number of errors detected by each group. N_{12} represents the number that both groups found in common. N is the total number of errors in the program, an unknown.

b. Assuming that all errors have the same chance of being detected, then the detection rate for a group applies not only to the entire space but also to a subset of that space.

$$D_1 = N_1/N = N_{12}/N_2$$

$$D_2 = N_2/N = N_{12}/N_1$$

or

$$N = N_{12}/(D_1 \times D_2)$$

As an illustration: if $N_1 = 20$ and $N_2 = 30$ and $N_{12} = 8$ errors, $N = 8/(8/30 \times 8/20) = 75$ total original number of errors.

c. Mr. Myers on page 336 also describes the error-seeding method for determining total error. This is also called the recapture method or the tagging method so popular in TV wildlife series. The method involves inserting randomly s errors into the program and then testing the program for error. When n is the number of detected indigenous errors and v is the number of detected seeded errors, then

$$N = sn/v$$

d. These methods may be useful in early stages of software development. However, in later stages these methods would lead to problems. Seeding errors in a large scale complex program running in real time can create havoc. The unexpected side effects to any change makes an integrated program best to leave alone. Using two groups and determining the faults found in common also is difficult in real time. While not deemed useful in software reliability, the methods may be useful in other areas in determining totals. One that comes to mind is determining total number of mines in a field.

PAGE 25-2, REVERSE, BLANK

Section 26

Use of Simulation to Reduce the Amount of At-Sea Testing

2601. Introduction. This paper psuedo-formalizes the joint use of simulation and at-sea testing to reduce the amount of at-sea testing.

a. The following illustrates the situation of possible use.

(1) We need to determine operational performance, using hit/miss type.

(2) This determination must be made by a certain time (deadline), or to a given confidence, or to a fixed cost.

(3) A small (insufficient) set of runs, say about 30, are made at sea.

(4) A simulation is available. The time/cost/effort in making a run on the simulation is deemed relatively small compared to that at sea. Typical ratios range from 1/100 to 1/700.

b. The procedure is as follows:

(1) The runs made at sea are duplicated on the simulation.

(2) The degree of correlation between results at sea and on the simulator is determined.

(3) Analysis should attempt to explain non-correlation effects. If this is successful, the correlation is effectively improved. If unsuccessful, the procedure would be invalid unless the non-correlation factors are deemed random. (This is an important limitation and more work has to be done.)

(4) If deemed applicable, make additional runs on the simulator. Use the formulae given in paragraph 2602 to adjust for bias, determine sample size, etc.

(5) If cost is important, use the factor in Table 26-1 to determine efficiency of making more runs.

c. The formulae are based on References a and b. It is important to note that these references do not pertain to simulation; 02B has projected its use to simulations. Before actual use of this approach, the references should be read.

2602. Illustration. This example is specific, giving the symbols, formulae, and an example count, using SIM to represent the simula-

tion and SEA to represent the at-sea results, which are considered infallible.

a. The first step consisted of 30 runs at SEA being duplicated on the SIM. The results were:

		SIM		
		Miss	Hit	Both
SEA	Miss	12	5	17
	Hit	2	11	13
	Both	14	16	30

b. Using the formulae in Ref a:

(1) Based on SEA alone, the hit probability (p) = 13/30 = 0.43.

(2) Based on SIM alone, the probability (π) = 16/30 = 0.53.

(3) The bias (optimistic in this case) is $p-\pi$ = -0.10.

(4) When SEA "hits", the misclassification probability (Θ) is 2/13 = 0.15

(5) When SEA "misses", the misclassification probability (Φ) is 5/17 = 0.29.

c. The correlation index $(r^2) = \frac{p(1-p)\ (1-\Theta-\Phi)^2}{\pi\ (1-\pi)} = 0.31.$

(1) This r^2 measures how well SEA data can be predicted from a run on SIM. For our purposes it can be interpreted as 31% of the factors influencing the run result at SEA are taken into account by the SIM. This means that 69% of the factors are not taken into account; if these factors are assumed to be random, i.e., merely contributing to the experimental error at sea, then the procedure is not invalid. Another way to put this: If the project objective is merely to estimate a single, common p, and only one scenario is involved in our at-sea testing, then the procedure is valid even with a low correlation.

(2) When $r^2 \rightarrow 0$, SIM results are independent of SEA results and use of SIM is wasteful. When $r^2 \rightarrow 1.0$, further runs at

SEA are wasteful. The observed value of 0.31 is considered poor, but some information is available from each run. A large number of runs on the SIM may have to be made to compensate for the poor correlation. Depending on the relative cost, this may still be worthwhile.

d. Suppose 384 additional runs are made on the SIM with 210 runs (55%) leading to hits. This and the original data base and symbols are:

		SIM		
		Miss	Hit	Both
SEA	Miss	n_{00} 12	n_{01} 5	$n_{0.}$ 17
	Hit	n_{10} 2	n_{11} 11	$n_{1.}$ 13
	Both	$n_{.0}$ 14	$n_{.1}$ 16	n 30

Y=174 X=210 N-n=384

e. Using all of the information available, the new estimate of hit probability at SEA is:

$$\hat{p} = \frac{n_{11}}{n_{.1}} \cdot \frac{n_{.1} + X}{N} + \frac{n_{10}}{n_{.0}} \cdot \frac{n_{.0} + Y}{N}$$

which is the sum of weighting the percentage hit on the SIM with its correct hit classification capability and the percentage missed on the SIM with its miss misclassification capability. In this example:

$$\hat{p} = \frac{11}{16} \cdot \frac{16 + 210}{414} + \frac{2}{14} \cdot \frac{14 + 174}{414} = 0.44$$

this 0.44 compares to the 0.43 obtained without the added information in the SIM runs.

f. The variance of this new estimate is:

$$V = \frac{p(1-p)}{n}(-r^2) + \frac{P(1-p)}{N}(r^2) =$$

$$\frac{(.44)(.56)}{30}(.69) + \frac{(.44)(.56)}{414}(.31) = 0.0059$$

This compares to the original estimate based only on the 30 runs at SEA:

$$V = \frac{(p)(1-p)}{n} = \frac{(.43)(.57)}{30} = 0.0082$$

g. Benefit. The change in hit probability is minor in this illustration. The change in variance is important and illustrates its value. This reduction in variance (0.0082 to 0.0059) based on 414 runs on SIM is similar to the reduction in increasing the runs at SEA from 30 to 42 on a confidence basis.

2603. Adaption To Our Work

a. There is no question but that this procedure is of limited use in our work. Its use may be more "in desperation" when the amount of at-sea services is so insufficient that reinforcement of results by other means is needed.

b. The procedure points out two things of interest. Runs on the simulator, if correlation is high, do supply some information that should be extracted. Runs on the simulator compared to those at sea are usually an order of magnitude cheaper in effort, time, and funding.

c. Use in our work depends on the random nature of the factors leading to low correlation; the broader the at-sea testing (different types of scenarios), the more critical this randomness assumption becomes.

d. Prior to determining correlation, the matched pairs/mismatched pairs should be analyzed for possible clues as to cause of mismatch. Sensitivity studies may help. If causes are found, the situation may be straightforward. For example the simulation may not account for reverberation. The procedure can be limited to non-reverberation conditions at sea. Most likely this search for causes will be fruitless. Most likely the simulation will be optimistic, the mismatches can be explained by usual (but not predictable) occurrences at sea. Then the question is, can the assumption be made that the observed correlation is general and not related to particular conditions? This implies that we can use as standard the correct hit classification ($n_{11}/n.1$) and incorrect misclassification ability ($n_{10}/n.0$). These two weights can then be used to determine p at each condition tested on the simulator.

e. For example, we may test 64 different test conditions, say in matrix form. For each condition we may make 25 runs. (Total number of runs is 1600, equivalent cost-wise to perhaps three runs at sea.) The 25 runs (number of hits/misses) at each condition can then be adjusted to the number of at-sea hits by

use of the standard weights. Analysis can then be made of these at-sea hits with the results pertaining to at-sea results.

2604. Use as Diagnosis

a. Parzen (see reference c) has an example which is pertinent to the use of this technique. Suppose we have the following (following Parzen):

		Simulation	
		Miss	Hit
S E A	Miss	.95	.05
	Hit	.05	.95

We would have a highly reliable simulation: it correctly predicts when SEA hits, 95% of the time. Now, if the SEA hit rate is real low, to make the point obvious, say 0.005, then by use of Bayes Theorem we get:

		SIM		
		Miss	Hit	Both
S	Miss	.94525	.04975	.995
E	Hit	.00025	.00475	.005
A	Both	.9455	.0545	1.00

Now .00475/.0545 = 0.09. This means that in 9% of the cases when SIM says hit, SEA will actually be a hit. At first glance this 9% is a contradiction to the 95% implied earlier. On reflection, though, with such a low success rate (.005), the misclassification (1-.95 = 0.05), while low per se, applies to most of the runs (0.995), which gives it a high weight.

b. Note that for the above case when SIM says miss, it is correct 99.97% of the time. (Frankly with such a low hit rate at sea, this is of no real importance.) When the SEA hit rate is .5 vice .005, we obtain 95% vice 9%. Thus, the original table applies (prior to using Bayes Theorem). When the misclassification rates are different for hits and misses, then the original (prior to Bayes Theorem use) classification rates hold only for the particular SEA hit rate present when these classification rates were determined. In other words Bayes Theorem must be used in conjunction with the SEA/SIM rates when applied to conditions other than tested. This in effect limits this technique to the conditions tested. If we knew the SEA hit rate to apply Bayes Theorem, we wouldn't have to use SIM at all.

Table 26-1

Percent Reduction in Variance by Use of Simulation

Correlation r^2	Relative Cost 25	50	100
30	10	16	20
50	25	35	40
70	49	56	60
90	74	80	83

Note: Based on a linear cost model. See Reference (a)

2605. References

a. Tenenbeim, A. A double sampling scheme for estimating from binominal data with misclassifications: Sample size determination. Biometrics, 27, No. 5, 935-44, Dec 1971.

b. Tenenbeim, A. A double sampling scheme for estimating from misclassified multinominal data with applications to sampling inspections. Technometrics, 14 No. 1, 187-202,Feb 1972.

c. Parzen, Modern Probability Theory and its Applications, John Wiley & Sons, (pp 119-120.)

Section 27

Performance Testing With Insufficient Sample Size

2701. General Comments. In many projects we find ourselves with an insufficient sample size. While this is to be deplored and avoided, it is often the case; we may have to "OPEVAL" with, say, only eight firings. This paper concerns this type of situation.

a. In this type of situation, the analyst can insure being less wrong, doing least harm with his techniques than without them. His "test design" may not be valid, but it still may be much better than none-at-all. The report may not be valid in a textbook fashion, but it still may materially aid the decision-makers. The point is that in an extremely tight situation, numerous judgements must be made; the analyst is in the best position to make many of these.

b. Prior to the actual firings there should be a regrouping of the scope, structure, objectives, test design, etc. of the evaluation. Sensitivity studies may be made with computer simulations. Rehearsal of planned firings including likely variations about the planned conditions should be made. The simulations may help in determining the operational impact of combining dry run results with firing results.

c. Each firing should have as complete instrumentation as possible, not only in all functions but also in the engineering as well as the operational areas. Each firing should be examined for deterministic results based on engineering experience.

d. The instrumenation should be of continuous-type data (say, 6.3' miss distance) vice merely hit/miss type.

e. The analyst should use test designs in selection of test conditions and/or to handle different environments, etc.

f. A functional approach is necessary. Each function (detection, classification, acquisition, homing, fuzing) should be instrumented and measurements made. This is important for its own sake as an analysis procedure. It is necessary to permit use of information from dry runs, and other data sources.

2702. Selecting Test Conditions

a. This illustrates using test designs for selecting test conditions. For impact an extremely efficient scheme is used, based on a few stringent assumptions. One critical assumption is that some key scenarios are of primary interest; these can be structured into test variables; these in turn can be integrated into one overall factorial scheme. Naturally the key scenarios

are included as cells in the factorial matrix. However, they need not be actually tested. Another assumption is that all interactions are of secondary importance in affecting results.

b. The illustration concerns seven test variables integrated into a factorial test plan. See Figure 27-1. Each variable is at two settings. The bar over a variable follows convention in indicating the absence of the variable. The matrix in Figure 27-1 displays 128 possible test conditions. A few of these may be scenarios of prim ry interest; many may be of secondary interest.

c. There is a total of eight valid firings or test runs in the evaluation. One possible set of eight (following Yates) is indicated in Figure 27-1. Notice the complete balance; there are four firings at each of the two settings for each of the seven variables.

d. A heuristic "proof" is used to illustrate the validity. For illustration purposes the example is free of experimental error. The example covers four phases.

(1) Arbitrarily setting true values which are the "effects" of the test variables. While arbitrary, the values are of various types. Lower case letters stand for these effects in Table 27-1.

(2) Delineating the run conditions and the resultant test data. See Table 27-2. The conditions are derived directly from the eight indicated in the matrix. The data are derived from the true effects in Table 27-1. For example, Run 4 is with variables A, F, and G at high setting. The datum, 217 units, is the sum:

5 for $\underline{A}$
30 for $\underline{B}$
100 for $\underline{C}$
10 for $\underline{D}$
12 for $\underline{E}$
20 for F
40 for G

(3) The analysis is straightforward since no interactions are assumed. The analysis is shown in Table 27-3, the resultant effects are also shown in this table. In a real test situation with experimental error, significance testing would be necessary using a guesstimate of the error based on experience. In this illustration, the determined effects agree with the true effects built into the illustration. This indicates validity (orthogonality) of the plan (within the stated assumptions).

Raid Type	Off Board Decoys	Defense Type	Missile Type	Clear				SOJ			
				$\overline{\text{RGPO}}$		RGPO		$\overline{\text{RGPO}}$		RGPO	
				$\overline{\text{land}}$	land	$\overline{\text{land}}$	land	$\overline{\text{land}}$	land	$\overline{\text{land}}$	land
Sparse	$\overline{\text{Chaff}}$	Self	$\overline{G}$								
			G								
		Mutual	$\overline{G}$				5				
			G					4			
	Chaff	Self	$\overline{G}$						3		
			G			6					
		Mutual	$\overline{G}$								
			G								
Dense	$\overline{\text{Chaff}}$	Self	$\overline{G}$							2	
			G		7						
		Mutual	$\overline{G}$								
			G								
	Chaff	Self	$\overline{G}$								
			G								
		Mutual	$\overline{G}$	8							
			G								1

Figure 27-1

Factorial Illustration of 2^7 showing Selected 8 Runs

(4) These effects are then used to predict the results at each scenario or cell in the matrix of Figure 27-1. For example the first cell value is 152, i.e., the effects of each variable at the absent (or low) setting is 152. Each of the 128 scenario or cell effects can be so determined.

Table 27-1

The True Effect

Variable	Settings Low	High	Effect
A (SOJ)	0	5	a = 5
B (RGPO)	30	12	b = -18
C (Land)	100	105	c = 5
D (Raid)	10	10	d = 0
E (Decoy)	12	30	e = 18
F (Defense)	0	20	f = 20
G (Missile)	0	40	g = 40

Table 27-2

Run Conditions and Test Data

Run Number	Variable Settings A	B	C	D	E	F	G	Data
1	H	H	H	H	H	H	H	222
2	H	H	L	H	L	L	L	139
3	H	L	H	L	H	L	L	180
4	H	L	L	L	L	H	H	217
5	L	H	H	L	L	H	L	159
6	L	H	L	L	H	L	H	192
7	L	L	H	H	L	L	H	197
8	L	L	L	H	H	H	L	190

Table 27-3

Analysis

4a = Runs 1,2,3,4 - Runs 5,6,7,8 = 758 - 738	a = 5
4b = Runs 1,2,5,6 - Runs 3,4,7,8 = 712 - 784	b = -18
4c = Runs 1,3,5,7 - Runs 2,4,6,8 = 758 - 738	c = 5
4d = Runs 1,2,7,8 - Runs 3,4,5,6 = 748 - 748	d = 0
4e = Runs 1,3,6,8 - Runs 2,4,5,7 = 784 - 712	e = 18
4f = Runs 1,4,5,8 - Runs 2,3,6,7 = 788 - 708	f = 20
4g = Runs 1,4,6,7 - Runs 2,3,5,8 = 828 - 668	g = 40

2703. Handling Systematic Change

a. Usually the key effort in experimental design is directed to integrating various scenarios into some factorial type. Then each trial furnishes information on all scenarios simultaneously. This leads to high precision and savings in sample size. Another feature of test design is directed to handling systematic changes in extraneous variables in our testing. Crew practice is a typical case in point. Other things being equal, the effect of practice on test data is less for runs made close together in time than for runs made far apart in time. Ship-to-ship differences are another example. Runs made, same ship, are more alike than ship-to-ship. In test design jargon this is called blocking. There are numerous block-type designs; Chain Block Design is one such type.

b. When scenarios cannot be related, when the number of scenarios is more than the block size, when flexibility in testing is needed, and when the experimental error is relatively small, then a chain block type may be useful. Illustration: six scenarios (A, B, C, D, E, and F) to be tested with three crews (I, II, III) with a total of 12 runs.

Crew I	Crew II	Crew III
A	C	E
B	D	F
C	E	A
D	F	B

Notice that the scenarios are paired off, linking various crews. Scenarios C and D link Crew I and II. Scenarios E and F link Crews II with III. And Scenarios A and B link Crews III and I.

c. The analysis first determines the block effect (which also may be of direct interest) and then "removes" these effects from the estimates of each scenario. The result is increased precision of the scenario estimates. The design is basically due to Youden.

Note: Experimental Designs by Cochran and Cox (Wiley, 1957) is an excellent source material. Cuthbert Daniel, "One-At-A-Time Plans" in JASA, June 1973 presents some small sample plans more in line with operational testing.

Section 28

Sample Size for MTBF

2801. Introduction. In determining if the system to be evaluated meets the MTBF threshold, we will form an acceptance testing plan of the following type. We will stress the system for a certain length of time. We will accept/reject depending on the number of failures observed. (Repair or replacement, exponential is assumed.) There is a choice of plans. For example, suppose the threshold criterion is MTBF $\geq$100 hrs. To demonstrate this with high confidence, say 80%,

a. We can test for 160 hours and accept if no failures.

b. We can test for 300 hours and accept with one failure.

c. We can test for 430 hours and accept with two failures.

d. Etc.

Thus, we can test for different numbers of hours (sample size). All at the same 80% confidence (false acceptance). Why not always select the testing scheme with the least test time? As the rest of this chapter points out, this is poor practice. If we use the minimum time with a no-failure scheme, we increase the risk of rejecting good systems. This is false rejection. Includes forcing the producer to "gold-plate" his system to insure passing our testing program.

2802. Symbols and Definitions

a. $L(\theta)$ is the probability of acceptance when the MTBF is θ.

$$L(\theta) = \sum_{k=0}^{c} \frac{e^{-Y}(Y)^k}{k!}$$

Where $Y = T/\theta$

b. θ_1 is slightly less than the threshold value MTBF. A system with true θ_1 is a POOR system. (This is the lower MTFB in MIL-STD-781C.)

c. θ_0 is the producer's quality relevant to the false rejection risk. A system with true θ_0 is a GOOD system.

d. α is the false rejection risk. It is the probability of rejecting equipment with a true MTFB equal to θ_0. Also called the producer's risk. $1-\alpha$ is the probability of acceptance.

e. β is the false acceptance risk. It is the probability of accepting equipment with a true MTBF slightly less than the threshold MTBF (θ_1). Also called the consumer's risk.

f. T is the total test time (sample size).

g. c is acceptance number; minimum number of failures allowed for acceptance.

2803. Acceptance Plan

a. Once the total test time and acceptance number are determined, the decision rule or acceptance plan is complete. The total test time and acceptance number are directly related to definition of POOR and GOOD and the corresponding acceptance and rejection risks. In other words T/c, θ_1, θ_0, α, and β are related. To determine one, the others must be fixed. To determine T/c, the others can be fixed as follows.

(1) θ_1, defined as POOR, is taken as the threshold value. (The "slightly less than" is ignored for all practical purposes.)

(2) β, defined as the Navy's risk, is taken as 0.20. This is based on OPTEVFOR policy of using 80% confidence limits with MTBFs.

(3) θ_0, defined as GOOD, could be taken as the goal value in the operational requirements. However, there must be a wide spread between threshold and goal for this to be feasible. If the ratio of θ_0/θ_1 is not large, say less than 2, than the procedure is not direct. See paragraph 2804.

(4) α, the false rejection risk, can be fixed arbitrarily. On one hand, this risk should not be less than the false acceptance risk (0.20). On the other hand, we should not be too stringent and have an excessive high "rejection of good systems" risk (>0.40). Except for unusual situations, α = 0.30 is a reasonable value.

b. Illustration: Suppose the MTBF threshold is 100 hours and the goal is 200 hours. Suppose also that project is typical, i.e, α = 0.30 and β = 0.20. With these values, we can go directly to Table 28-1. The entries of interest are 0.20 Navy's risk and the 0.70 acceptance column. In using these entries, first convert the test time factor (col. 3) and the producer's quality factor (col. 5) into hours by multiplying by the threshold value of 100 hours. This gives one entry that fits our criterion: producer's quality of 200 hours when we test for 550 hours and allow three failures. The sample size answer in this case is 550 hours. The decision acceptance rule is 550 hours with three failures. (This can be checked by calculating L(100) and L(200) using the L(θ)

formula given in paragraph 2802. L(200) = 0.70 and L(100) = 0.20.

c. Note that if the producer designs his system only to meet the threshold, the odds are high (0.80) it will be rejected. That is the meaning of L(100) = 0.20.

d. Note that not all the criteria values may fit exactly because of the discrete nature of the acceptance number. For example if the goal or producer's quality is 210 hours, then using 500 hours and three failures in the acceptance plan, the actual false rejection risk is 0.27 vice 0.30.

2804. Trade-offs. If, for various situations, the parameters cannot be determined in a straightforward fashion, then some trade-offs are necessary. θ_1 is still taken as the threshold value; β is still fixed at 0.20. The trade-off comes in selecting from Table 28-1 the one T/C plan of the type mentioned in 2801. Thus, we can test for 160 hours, accept if no failures. This will give us the 20% false acceptance risk. However, to have a reasonable false rejection risk, the producer's quality must be 3 to 7 times more than the threshold. Thus, in minimizing test time:

a. We increase the risk of rejecting systems which truly surpass the threshold, or

b. We may force the producer to gold-plate the system. We may have to pay for this gold plate for each installation over the next decade.

With respect to these trade-offs we should strive for test time with at least three failures in the acceptance plan. A review of Table 28-1 indicates this as a reasonable trade-off.

2805. Source. This chapter presents the OC (operating characteristics) curve approach to sample size. See Figure 28-1. Two papers directly applicable to MTBF are B, Epstein, "Statistical Techniques in Life Testing," Technical Report #3, ONR Contract Number 2163(00) (NR042-081) of June 1, 1959 and also MIL-STD-781C. Note that the familiar TEAM Reliability Slide Rule also can be used. The confidence level with this slide rule is associated with the consumer's risk as $1-\beta$.

2806. α vice β. The use of α and β in this chapter may appear to be opposite to the standard jargon, particularly for analysts trained in hypothesis testing. This is not surprising. The translation of α and β errors from hypothesis testing to acceptance testing is not direct. The relationship between consumer's and producer's risks and α and β is not self-evident. Also a source of confusion is the standard use of α for table look-up

even though the β error is the value in the table. Finally, the term "confidence" may pertain to a wide spectrum, α, β, 1-α, or none of the above.

a. In Hypothesis Testing, α (or Type I Error) is the probability of rejecting H_o (or the null hypothesis) when it is true. For example, if the H_o is p = 0.60 and the system has a true quality = 0.60, than α is the risk of our test rejecting H_o, p = 0.60, even though it is true. In Acceptance Sampling Testing, we define a certain level of quality as being GOOD. This is θ_o in the exponential case, or P_o in the binomial case (see Section 29). If the system has a true quality of P_o (or better), it is a GOOD system; α is the risk of rejecting a GOOD system (P_o) with our sampling plan. For example, if the definition of GOOD = P_o = 0.60, and the system is GOOD, P_o = 0.60, then α is the risk of our sampling plan rejecting the system as good even though it is GOOD. NOTE: The producer must produce a system as good as GOOD to insure its acceptance 1-α probability.

b. In Hypothesis Testing, β is the probability of accepting the H_o hypothesis when it is false. Of course, this β or Type II Error will vary with "how much a departure from H_o." This is the reason for the alternative hypothesis. For example, if the H_o is p = 0.60 and the system has a true value of p = 0.10, the β will be much less than if p = 0.59. In Acceptance Testing, the alternative hypothesis is the definition of POOR. In the exponential case it is θ_1, slightly less than the threshold. In the binomial it is θ, slightly less than the threshold. β than is the probability of accepting the system (as GOOD) when it is actually POOR. This is the consumer's risk.

2807. Note. Sections 29 and 30 should be scanned as they are related to this general subject. The total sample size discussion on page 11-5 may be pertinent.

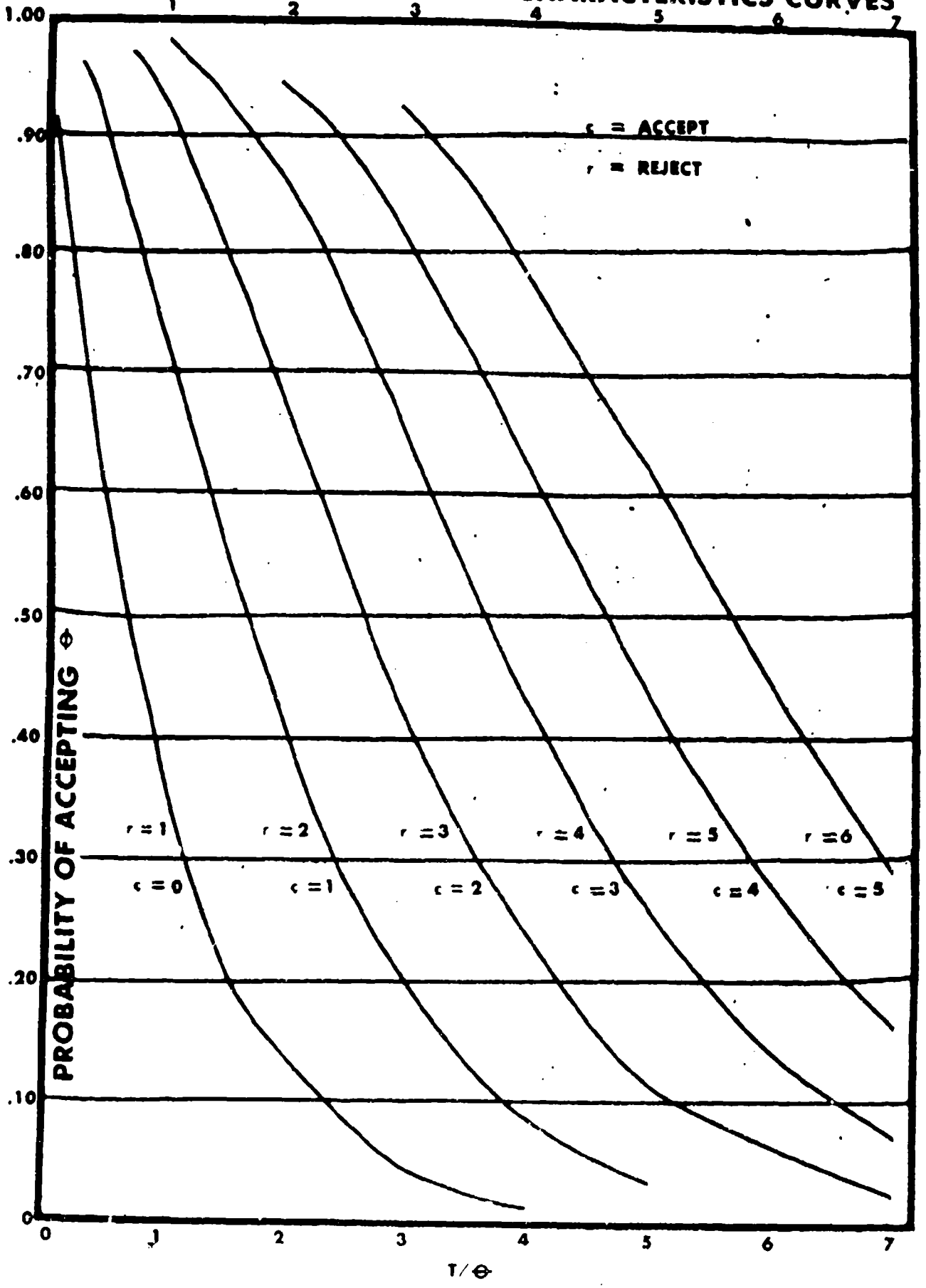
Table 28-1

MTBF Test Time Determinations* and Producer's Quality
All Values in Terms of Threshold MTBF

False Acceptance Risk (β)	Acceptable Number of Failures	Test Time*	Needed Producer's Quality* by False Rejection Risk (α) .20	.30	.40
.10	0	2.3	10.2	6.5	4.5
	1	3.8	4.7	3.5	2.8
	2	5.2	3.4	2.7	2.3
	3	6.7	2.9	2.4	2.1
	4	8.0	2.6	2.2	1.9
	5	9.2	2.4	2.0	1.8
.20	0	1.6	7.2	4.4	3.2
	1	3.0	3.6	2.7	2.2
	2	4.3	2.8	2.3	1.9
	3	5.5	2.4	2.0	1.7
	4	6.7	2.2	1.8	1.6
	5	7.8	2.0	1.7	1.5
.30	0	1.2	5.4	3.4	2.4
	1	2.4	2.9	2.2	1.7
	2	3.6	2.4	1.9	1.6
	3	4.8	2.1	1.7	1.5
	4	5.8	1.9	1.6	1.4
	5	7.0	1.8	1.6	1.4

*Multiply by MTBF threshold time.

FIGURE 28.1 OPERATING CHARACTERISTICS CURVES



Section 29

Sample Size, Binomial

2901. Introduction. In determining if a system meets a binomial type threshold, we will form an acceptance sampling plan of the following type. We will stress the system a certain number of times or we will stress a certain number of items. There is a choice of plans. For example, if the threshold criterion is p = 0.85, to demonstrate this at the 10% consumer's risk or false acceptance risk:

a. We can test 15 items with no failures.

b. We can test 25 items with one failure.

c. We can test 36 items with two failures.

The differences among these plans are the different false rejection risks. Suppose we elect to test 36 items, allowing two failures. The producer, to insure, say, 90% confidence to have his system pass, has to produce a system whose success rate is 0.98. If we elect to test only 15 items, allowing no failures, the corresponding producer's quality must be 0.993. This is not only costly but increases the risk of rejecting good systems.

2902. Symbols and Definitions

a. L(p) is the probability of acceptance when the quality is p.

b. p_1 is slightly less than the threshold value.

c. p_0 is the producer's quality relevant to the false rejection risk.

d. α is the false rejection risk, the probability of rejecting equipment with a true quality equal to p_0. $1-\alpha$ is the probability of acceptance.

e. β is the false acceptance risk, the probability of accepting equipment with a true quality slightly below the threshold.

f. n is the sample size.

g. c is the acceptance number; minimum number of failures for acceptance.

2903. Binomial. The usual binomial expression is

$$L(p) = \sum_{x=0}^{c} \frac{n!}{x!(n-x)!} (1-p)^x p^{n-x}$$

which is used to determine the risk associated with various sampling plans. Note that the TEAM Reliability Slide Rule can help. The confidence level with this slide rule is associated with the false acceptance risk as $1-\beta$. The slide rule is useful in obtaining the various sampling plans directly for a selected β. Then the above binomial expression is used to obtain various producer's qualities and false rejection risks for trade-off in selecting the specific acceptance plan.

2904. Illustration of Use. Suppose the threshold success rate, p, is 0.85. Suppose also that being a reliability threshold, we have selected a consumer's risk, β, of 0.20. Using the TEAM Reliability Slide Rule with 80% confidence we get

$c = 0$, $n = 10$
$c = 1$, $n = 20$
$c = 2$, $n = 28$
etc.

The next step is to select one of these various possibilities. (Note that if the slide rule is not handy, we can use the binomial expression to get these values.) Then, a working table is derived, such as Table 29-1. With such information, various tradeoffs can be made. Perhaps, a typical selection would be testing 28 items with two failures. The important point is that both the risk factors and corresponding success rate should be given so that all parties concerned understand the specific risks.

2905. Note. Please read Section 28 for more explanation of this operating characteristics approach, use of goals, etc.

Table 29-1

Probability of Accepting (L(p)) Various p Values
(Work Table, $\beta = .20$, $p_1 = .85$)

p	Probability of Acceptance		
	$c_0 = 0$, $n_0 = 10$	$c_1 = 1$, $n_1 = 19$	$c_2 = 2$, $n_2 = 28$
.85	.20	.20	.19
.88	.28	.32	.33
.90	.35	.42	.46
.92	.43	.54	.61
.95	.60	.75	.84
.98	.82	.95	.98
.99	.90	.98	1.00

NOTE: When $c = 0$, $L(p) = p^{n_0}$

When $c = 1$, $L(p) = p^{n_1} + n_1(q)^1(p)^{n_1-1}$

When $c = 2$, $L(p) = p^{n_2} + n_2(q)^1(p)^{n_2-1} + \frac{1}{2}(n_2)(n_2-1)(q)^2(p)^{n_2-2}$

Note: Page 29-4 gives a program to obtain L(p) values for c = 0, 1 or 2 for a Hewlett-Packard 25 hand calculator.

Note: Binomial tables should be used for large c values.

Hewlett-Packard 25 Program to Calculate L(p) for C = 0, 1, 2 only

1. Certain items must be manually stored:

n in R_2, n-1 in R_3, n-2 in R_4, c in R_7. Then to begin, GTO 00, key in p, R/S and key in 1-p, R/S. Display gives L(p), cycle thru the next keying of p, etc.

Display Line	Code	Key Entry	Comments
00			
01	2300	STO 0	Stop p
02	74	R/S	
03	2301	STO 1	Store 1-p
04	02	2	
05	2305	STO 5	
06	2400	RCL 0	
07	2402	RCL 2	
08	1403	fy^x	
09	2306	STO 6	
10	2407	RCL 7	
11	1571	gx=0	c = 0
12	1339	GTO 39	
13	2400	RCL 0	
14	2403	RCL 3	
15	1403	fy^x	
16	2401	RCL 1	
17	61	x	
18	2402	RCL 2	
19	61	x	
20	235106	STO+6	

Display Line	Code	Key Entry	Comments
21	1574	NOP	
22	2405	RCL 5	
23	2407	RCL 7	
24	1441	fx<y	
25	1339	GTO 39	c = 1
26	2400	RCL0	
27	2404	RCL 4	
28	1403	fy^x	
29	2401	RCL1	
30	1502	gx^2	
31	61	x	
32	2403	RCL 3	
33	61	x	
34	2402	RCL 2	
35	61	x	
36	2405	RCL 5	
37	71	÷	
38	235106	STO+6	
39	2406	RCL 6	c = 2
40	1300		
41			

Section 30

Sample Size for Mean Values (Normal Distribution)

3001. Introduction

a. In determining if the system to be evaluated meets a mean value threshold, we will form an acceptance sampling plan. We will make a certain number of runs with the system, collect data, and determine the mean and standard deviation. An important part of the acceptance plan is making a statistical t test to determine if the observed mean is significantly different from the threshold. If significantly different, we will accept the system if it meets the threshold; we will reject it if it does not meet the threshold.

b. This paper is based on the assumption of a normal distribution, mean and standard deviation are unknown. The criterion is one-sided, either higher or lower than the threshold. There is a choice of sampling plans. We can make, say, 10, 20 or 30 runs using the appropriate t test for significance and acceptance value. The differences among these plans are the different consumer's and producer's risks.

3002. Symbols and Definitions

a. $L(\mu)$ is the probability of acceptance when the true mean value is μ.

b. μ_1 is the threshold value of the mean

c. μ_o is the producer's quality relevant to the false rejection risk α.

d. α is the producer's risk or false rejection risk, the probability of rejecting a system whose true mean is equal to μ_o. $1-\alpha$ is the probability of acceptance.

e. β is the consumer's risk or false acceptance risk, the probability of accepting a system whose mean value is just equal to the threshold value, μ_1.

f. s is the observed standard deviation.

g. n is the sample size.

h. d is the standardized difference relative to the standard deviation, $d = (\mu - \mu_1)/\sigma$.

i. σ is the true standard deviation.

3003. Illustration of the Underlying Procedure

a. Suppose the threshold is given as no less than 150 kyd. In acceptance sampling jargon, this (or slightly less than this value) is considered "poor." Suppose the estimate of the standard deviation is 12 kyd and that the plan includes significance testing, after the data are collected, at a significance level of 0.05. Suppose we are planning on a sample size of 10.

μ_1 = 150 kyd

s = 12 kyd

n = 10

t = 1.833 at 0.05 significance level at 9 degrees of freedom.

b. The first step is to determine the decision or acceptance value. This is found by using the t statistic (1.833) with level 0.05 and n-1 degrees of freedom

$t=1.833=\sqrt{n}(\mu-\mu_1)/s = !10(\mu-150)/12$

This works out to μ = 157 kyd. So after the data are collected, the observed mean must be greater than 157 kyd for system acceptance. (The actual value may differ based on the s estimate.) The second step is determining the likelihood of obtaining a value greater than 157 kyd. This, of course, depends on the true quality of the system. Suppose the true quality of the system is 160 kyd. The probability of accepting the system is determined by the standard t ratio.

$t=\sqrt{10}(160-157)/12 = 0.79$

The tabular probability up to this value is about 0.79. Thus the probability is 0.79 of accepting a system whose quality is 160 kyd when we use the acceptance plan above.

c. Going through the same manipulations, the probability is 0.90 when the quality is 162 kyd. See Figure 30-1.

3004. Generalizing the Procedure

a. The relationship to obtain the acceptance value, μ_a

$\sqrt{n}(\mu_a-\mu_1)/s = t_a$

b. The only unknown is μ_a for given n and acceptance significance level:

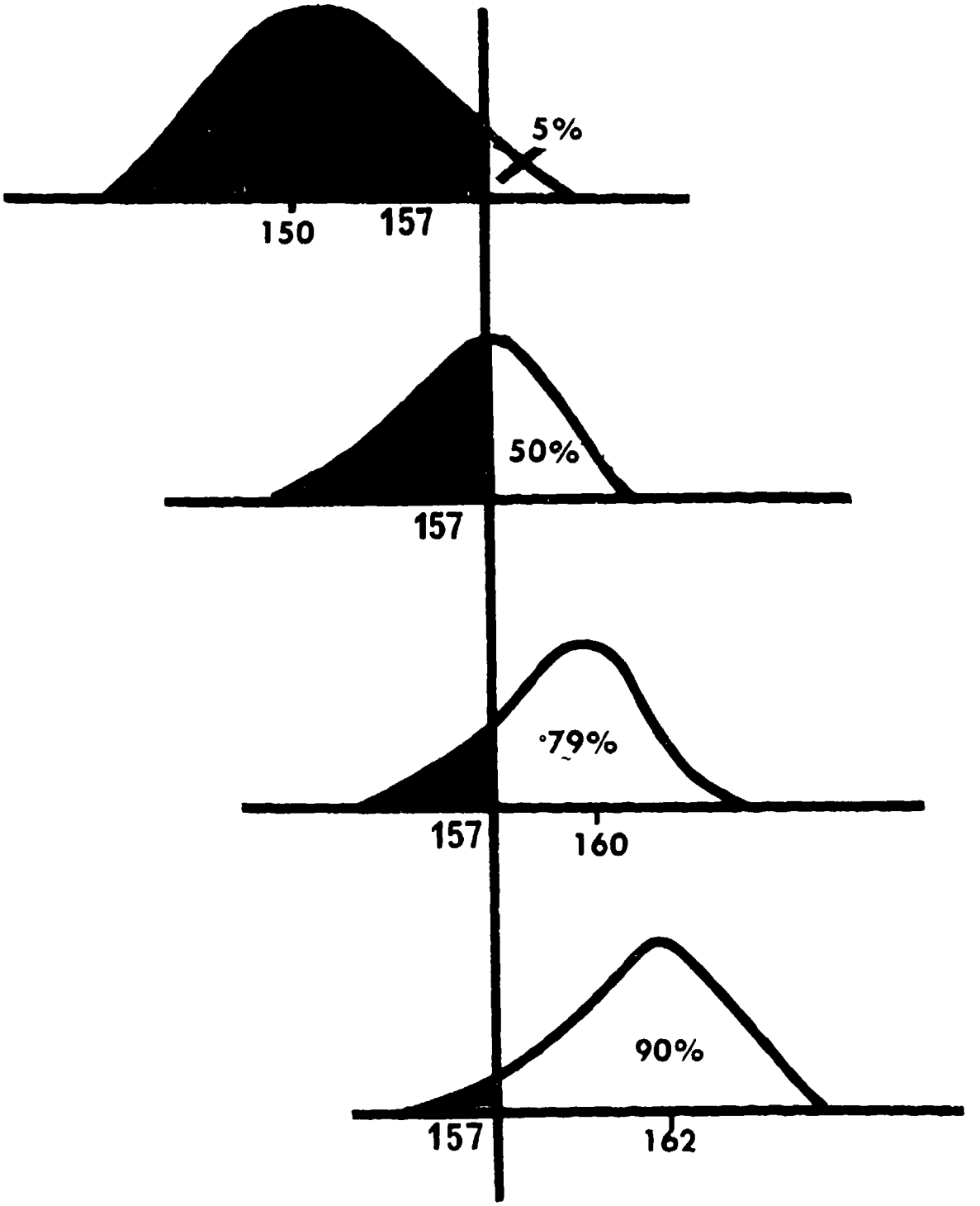


FIGURE 30-1
GRAPHICAL DISPLAY OFIP 3003
(NOT TO SCALE)

$$\mu_a = t_a \, s/\sqrt{n} + \mu_1$$

c. The t ratio is also used as:

$$t_o = \sqrt{n}(\mu_o - \mu_a)/s$$

d. Substituting for μ_a we get

$$t_o = \sqrt{n}[\mu_o/s - t_a/\sqrt{n} - \mu_1/s]$$

$$t_o = \sqrt{n}(\mu_o - \mu_1)/s - t_a$$

e. Based on how close σ is to s, then

$$t_o \sim \sqrt{n}\, d - t_a$$

f. Table 30-1 gives selected values. Note that these are approximate since the standard deviation is a guesstimate. This table is used as follows: The consumer's risk is selected. Based on the threshold value and a guesstimate of the standard deviation, the relationship between the producer's quality and producer's risk can be found for different sample sizes. Trade-off analysis can then be used to select the appropriate sample size. It is important to give the corresponding risks and quality levels involved so all concerned can understand the impact of the selected sample size.

3005. Note. Please read Section 28 for more explanation of this operating characteristics approach, trade-offs, use of goals, etc. Section 11 discusses sample size for comparisons and the total sample size concept.

Table 30-1

Probability of Acceptance as Meeting or Exceeding Threshold

d	Consumer's Risk											
	0.05				0.10				0.20			
	Sample Size				Sample Size				Sample Size			
	5	10	15	20	5	10	15	20	5	10	15	20
0.0	.05	.05	.05	.05	.10	.10	.10	.10	.20	.20	.20	.20
.2	.09	.12	.17	.22	.19	.25	.30	.35	.33	.40	.46	.50
.4	.14	.32	.41	.51	.28	.45	.57	.67	.50	.64	.74	.81
.6	.23	.52	.72	.82	.43	.70	.83	.90	.64	.83	.91	.95
.8	.37	.74	.89	.96	.60	.86	.95	-	.78	.93	-	-
1.0	.54	.90	.97	-	.74	.95	-	-	.86	-	-	-
1.2	.75	.96	-	-	.84	-	-	-	.92	-	-	-
1.4	.81	-	-	-	.91	-	-	-	.96	-	-	-
1.6	.88	-	-	-	.95	-	-	-	-	-	-	-

Section 3'

Hit Probability With Various Situations: Literature Listing

This chapter lists abstracts on the general area of determining hit probability against a target, usually circular. The coverage is mainly two-dimensional with either zero bias or a definite offset with usually single-shot. Results are usually exact. While the source of the article is given, the entire article may be obtained from 02B (Autovon 690-5177). 3114-3116 are exceptions. The abstracts are by the authors.

3101. THOMAS, M.A., and TAUB, A. E. "Salvo Kill Probabilities for Offset Aim Points", Naval Surface Weapons Center Report TR-3655, May 1977.

A mathematical model is formulated for calculating two error salvo kill probabilities when all rounds in the salvo are aimed at a point offset from the target center. The two errors associated with the warhead impact points are (1) an aiming or bias error assumed circular normal and common to all rounds in the salvo and (2) round-to-round or random errors also assumed circular normal which vary from round to round within the salvo. Extensive tables were prepared using numerical integration techniques and are provided as an aid to the weapons systems analyst. The program listings and input instructions are also provided to enable the analyst to compute probabilities for those cases which are not tabulated. Examples are provided which illustrate the use of the tables and program in obtaining individual salvo kill probabilities and in obtaining the expected number of targets killed when aiming at an intermediate point in a cluster of targets.

3102. DIDONATO, A. R., JARNAGIN, M. P., JR., HAGEMAN, R. K. "Kill Probability of a Gaussian Distributed Cookie-Cluster Weapon Against a Random Uniformly Distributed Point Target within an Ellipse", Naval Surface Weapons Center Report TR-3453, Nov. 1976.

A solution by deterministic methods is described of the problem of computing the single-shot kill probability of a point target at a random point from a uniform distribution over the interior of an arbitrary ellipse in the plane, given that the distribution of shots is uncorrelated bivariate normal with respect to a rectangular coordinate system in the plane, and that the weapon has a cookie-cutter damage function with prescribed lethal radius R. This solution has been programmed at NSWC, Dahlgren Laboratory. The numerical evaluation of a double integral whose integrand contains the so-called elloptic coverage function is required. Computer results clearly show that superiority of this solution over a non-deterministic, Monte Carlo method of Weidman and Brunner.

3103. THOMAS, Marlin A., "Salvo Kill Probabilities for Circular Targets - Axisymmetric Case", NWL TR-2643, Nov. 1971.

The kill probability resulting from the delivery of a salvo of weapons is not a straightforward calculation since the aiming error is likely to be common to all rounds in the salvo. In the absence of a computer program or a set of tables the analyst may have to resort to the binomial law $1-(1-p)^N$, where p, the "single shot kill probability" is computed on the assumption that both aiming error and round-to-round error vary with each round in the salvo. Use of the binomial law in the salvo case can introduce serious error.

As an aid in determining weapon requirements, comparing weapon systems effectiveness, etc., salvo kill probabilities against circular targets are tabulated for a wide variety of parametric values under the following assumptions: (1) one aims at the center of a target of radius a and fires a salvo of size N; (2) the error in the mean impact point of the salvo from the target center is governed by a circular normal density with variance σ_1^2; (3) the mean impact point is common to all rounds in the salvo but varies from salvo to salvo; (4) the errors in the individual impact point of shots within a salvo from the mean impact point are independently governed by a circular normal density with variance σ_2^2; (5) the two errors above, referred to as the aiming error and the round-to-round error, respectively, are independent. The salvo kill probability, i.e., the probability that at least one round in the salvo falls within the target, is computed as a function of $R = a/\sigma_2$, $T = \sigma_1/\sigma_2$, and N for R = .1(.1)3.0(.2)5.0, T = .1(.1)3.0(.2)5.0 and N = 1(1)14(2)20. Various examples pertaining to the use of the tables are given.

3104. McNOLTY, Frank, "Kill Probability for Multiple Shots" Operations Research, Vol. 15, No. 1, pp 165 169 1967.

A point target is randomly located according to an offset circular normal distribution and remains in its unknown position throughout N independent tosses of a lethal circle. The paper presents integral expressions for the probability of: (1) killing the target at least once in N tosses of the lethal circle; (2) killing the target exactly n times in N tosses; (3) requiring less than or equal to m shots to kill the target exactly once; (4) killing the target at least once in N tosses when the bias (offset distance) is randomly distributed. The paper also presents formulas for the expected number of shots required to kill the target exactly once and the expected number of times the target is killed in N tosses. A simple expression is also given for the single-shot kill probability for an offset ellipsoidal case when the lethal radius of the weapon is variable rather than fixed.

3105. CLODIUS, Fedric C., "A Model to Determine Kill Probabilities for a Salvo Weapon", NWL TR-R-18/66, April 1966.

This report describes a computer model that calculates a salvo-fire weapon's kill probabilities as a function of the number of elements in a salvo per burst, the aiming error associated with the weapon, the salvo dispersion, the target casualty criteria, and the target's location. Although this presentation applies specifically to personnel targets, the model is applicable for other target types where the conditional kill probability is known.

3106. McNOLTY, Frank, "Kill Probability when the Lethal Effect is Variable", Operations Research 13, 478-482(1965)

Traditionally the probability of killing a point target is analyzed by means of employing a lethal circle (in two dimensions) or a lethal sphere (in three dimensions). The concept of a lethal circle or sphere based on the assumption of a 'kill-no kill' situation in which a target anywhere within the circle (sphere) is killed and a target anywhere outside of the lethal region is not killed. Another aspect of the "kill-no kill' approach is that it combines all lethal effects (heat, shock, x-rays, etc.) in one parameter--the radius R of the lethal circle (sphere). In many problems of weapon systems effectiveness this approach is extremely useful, but in other areas the attendant assumptions are unrealistic and the purpose of this paper is to present a useful alternative methodology. All of the formulas derived are amenable to desk-calculator computation.

3107. JARNAGIN, M. P., Jr., "Expected Coverage of a Circular Target by Bombs All Aimed at the Center", NWL Report, No. 1941, June 30, 1965. DDC #AD-618-877.

The exact mathematical formulation is derived for the expected proportional coverage of a circular target from n weapons all aimed at the center of the target. It is assumed that the weapons fall in a circular normal distribution and that the lethal area of each weapon is a circle of specified radius (cookie-cutter lethality function). Tables are presented which are based on this formulation and which were computed on an IBM 7030 (STRETCH) computer. The computing time per case for coverages correct to two decimal digits is 0.25 to 0.50 seconds.

3108. MARSAGLIA, George, "Some Problems Involving Circular and Spherical Targets", Operations Research, pp 18-27, 1965.

This article is concerned with some problems that occur in certain tactical considerations: how should one place k circles (spheres) in the plane (3-space) so that their union has the greatest standard normal probability measure, that is, so as to maximize

the probability that a random normal point will fall in one or more of the circles (spheres). For $k>3$ the problem seems hopeless, (except for certain special situations); the case for $k=3$ is still unresolved and is being worked on by a number of investigators, and the case for $k=2$ is solved completely in this paper. The results for $k=2$ have some practical value when applied to actual problems arising in tactical considerations, and some theoretical value, as a method of attacking the problem for $k\geq 3$.

3109. GRUBBS, Frank E., "Approximate Circular and Noncircular Offset Probabilities of Hitting", Operations Research, pp 55-62, 1964.

For equal or unequal delivery errors and an offset point of aim, the chance that the burst point of a warhead occurs within a given distance of a selected point of the target is approximated by reference to weighted noncentral chi-square distributions. Offset circular and noncircular probabilities of hitting for the two and three dimensional cases may thus be approximated with a single, straight-forward and rather simple technique by the use of an approximate central chi-square distribution with fractional number of degrees of freedom or a transformation to approximate normality. Computations of probabilities of hitting are illustrated by examples. The approximations recommended appear to be of sufficient accuracy for many weapon systems evaluation problems. We do not claim originality for the various parts of theory involved, but rather our purpose is to provide the weapon systems analyst with a reasonable and useful analytical procedure.

3110. GILLILAND, Dennis C., "Integral of the Bivariate Normal Distribution over an Offset Circle", Journal American Statistical Association, Dec. 1962.

In problems where guidance to the proximity of a point is required, the probable success of the mission often is described by the probability of hitting within a given radius of the point. In this paper is presented a method of determining the probability of a hit within a given circle when it is assumed that the guidance error is distributed according to a bivariate normal distribution. This problem can be solved readily if the random error variables are independently distributed along two orthogonal axes with equal standard deviations and if they are not biased relative to the target location. Since these assumptions often are not valid, a series solution to the general problem is provided with an analysis of the error introduced by considering only a finite number of the terms.

3111. HILLIER, Ann, "A Program for Computing Probabilities over Rectangular Regions Under the Multivariate Normal Distribution", TR-54, 28 July 1961, Applied Mathematics and Statistics Laboratories, Standford Univ.

In recent years tables of the bivariate normal distribution and a table for computing trivariate normal probabilities have become available. Unfortunately, tabulation beyond three dimensions is rather unwieldy so that there is little hope of extensive tables being produced. However, there are many important applications which involve computing probabilities of multivariate normally distributed random variables beyond the dimension three. For example, suppose a complex item consists of ten dependent components. Suppose further that their measurable characteristics are random variables having a multivariate normal distribution with known vector of means and known variance covariance matrix. An upper limit for each measurable characteristic is given, and the complex item will fail if any of the measurable characteristics exceeds its upper limit. Calculating the probability of failure requires a knowledge of the ten-variate normal distribution function.

A second example deals with the tracking of a missile after firing. The actual path of the missile is a time series which can be measured at a finite number of points, say nine. Suppose that the projection of the path on the horizontal plane is the important variable; i.e., deviation from a fixed line is deemed critical, and the missile will be destroyed if the maximum deviation is too far from its desired path. If the deviations are assumed to have a multivariate normal distribution with known vector of means and known variance covariance matrix, the probability of destroying the missile can be computed from a tabulation of the nine-variate normal distribution.

Since a complete tabulation of the multivariate normal distribution is not feasible, a machine program which generates the required probabilities would be useful. It is the purpose of this technical report to present such a program. This program may be used to obtain probabilities over regions which are constructed by placing upper bounds on each of the variables. Thus, if $X_1, X_2, \ldots, X_n$ are multivariate normal random variables, and $U_1, U_2, \ldots, U_n$ are fixed numbers, the proposed program can be used to find

$$P(X_1 \leq U_1, X_2, \ldots, X_n \leq U_n) .$$

3112. HARTER, H. Leon, "Circular Error Proabilities", Journal of the American Statistical Association, Vo. 55(1960), December, pp. 723-731.

A problem which often arises in connection with the determination of probabilities of various miss distances of bombs and missiles is the following: Let x and y be two normally and independently distributed orthogonal components of the miss distance, each with mean zero and with standard deviations σ_x and σ_y, respectively, where for convenience one labels the components so that

$\sigma_x \geq \sigma_y$. Now for various values of $c = \sigma_y/\sigma_x$, it is required to determine (1) the probability P that the point of impact lies inside a circle with center at the target and radius $K\sigma_x$, and (2) the value of K such that the probability is P that the point of impact lies inside such a circle. Solutions of (1), for c = 0.0(0.1)1.0 and K = 0.1(0.1)5.8, and (2) for the same values of c and P = 0.5, 0.75, 0.9, 0.95, 0.975, 0.99, 0.995, 0.9975, and 0.999, are given along with some hypothetical examples of the application of the tables.

3113. JARNAGIN, M.P., and DiDONATO, A.P., "Damage to a Circular Target by a Gaussian Distributed Warhead with Uniformly Distributed Bomblets", Operations Research, 14, Nov-Dec 1966, pp. 1014-1023.

A derivation is given for an analytic expression representing the expected damage to a circular target of uniform value by a cluster of N bomblets. These bomblets are scattered by the detonation of a larger bomb. Each is considered to have a circular lethal area of specified radius, the same radius for all bomblets. The impact point of the large bomb is governed by a circular normal distribution with its mean at the target center. The bomblets are assumed to be independently dispersed under a uniform distribution over a prespecified circular area, known as the warhead-effects field, having its center at the impact point of the large bomb. Typical graphical results are included.

3114. PFEILATICKER, R., GLYNN, J. "Hit Probability on a Tank Type Target" Frankford Arsenal March 1966. DDC #AD-639-019.

A mathematical model was developed that may be used to determine salvo hit probability as a function of the number of rounds in the salvo, target size, aim error and round to round variation (dispersion). The report indicates graphically the value of dispersion that maximizes the salvo hit probability on a tank type target for a given number of rounds and a given aim error. The conclusions indicate that the dispersion should be approximately equal to the bias in order to maximize the salvo hit probability on a tank type target (7½ ft. x 7½ ft.). Although the model was developed based on the salvo assumption, it does serve to provide a measure of the effectiveness of high rate of fire weapons, where the relationship between bias and dispersion may approach that of a salvo. The salvo assumption, i.e. that the aim error (bias) is considered constant during the firing interval and round to round dispersion may be superimposed upon this bias, leads to hit probabilities less than those obtained by the 'independence of rounds' assumption generally employed for calculation of low rate of fire weapons. For high rate of fire weapons, actual hit probabilities would be expected to fall somewhere between the two solutions depending on the degree of round to round correlation existing in the fire control system.

3115. BISER, E., and MILLMAN, G., "Tables of Offset Circle Probabilities for a Normal Bivariate Elliptical Distribution" Army Electronics Command, Fort Monmouth, N.J. Aug. 1965 DDC #AD-623-882.

This report consists of two major parts. The first deals with the development of formulas for computing the probability that a point taken from a normal bivariate elliptical distribution with specified mean and standard deviations shall fall within a circle of given radius whose center is displaced a given distance from the center of the distribution. The second part consists entirely of probability tables. These tables will prove especially useful in dealing with problems involving accuracy studies of weapons systems and with other problems notably in meteorological studies. The events in many practical probability problems are best described by a normal bivariate elliptical distribution with unequal standard deviations. For example, one may be confronted with the problem of evaluating the probability that a missile will hit a circle of a specified radius whose center (aim point) is displaced a given distance from the mean (of impact points) of a normal bivariate elliptical distribution. In this example the impact points are governed by a normal (Gaussian) bivariate elliptical density function; the mean of this distribution is not zero (i.e., the center of the distribution is not about the aim point).

3116. Thomas, M., J. Crigler, G. Gemmill, and A. Taub, 1973. Tolerance Limits for the Rayleigh (Radial Normal) Distribution with Emphasis on the CEP, NWL Technical Report TR-2946, Naval Weapons Laboratory, May 1973.

The weapon systems analyst is often confronted with the problem of estimating the radius of a mean centered circle which will include 50% of the future rounds from a particular weapon under specified conditions. If the fall of shot tends to follow a circular normal distribution with standard deviation σ, this is usually accomplished by estimating σ with σ' from the results of test firings and then forming $\widehat{CEP}' = 1.774\hat{\sigma}'$. CEP, the parameter estimated, is the radius of a mean centered circle which includes 50% of the bivariate probability which, of course, is taken to mean 50% of the future rounds from this weapon under similar conditions.

While CEP′ is a valid point estimate of the radius of the 50% circle (provided σ' is a valid estimate of σ), it does not provide the analyst with any measure of confidence concerning his statement.

Statements of confidence concerning the percent of a population which lies within a circle of given radius are formulated in this report through the concept of statistical tolerance limits. The results will enable the analyst to ascertain (with the aid of tables) the confidence with which he can state that a circle of radius CEP′ contains at least 50% of the population. This confidence is shown to be quite low (at most .50 unless one has complete knowledge about the population parameter σ) and can be increased

only by increasing the multiplying constant for σ' above the customary 1.774. Tables of such constants (tolerance limit factors) are provided which will enable the analyst to obtain more reasonable levels of confidence not only for 50% of the population but also for 75%, 90%, 95%, and 99%.

3117. Thomas, M.A. and Crigler, J.R. and Gemmill, G.W., and Taub, A.E. "Tolerance Limits for the Maxwell Distribution with Emphasis on the SEP" Navy Weapons Laboratory TR-2954, June 1973.

Tolerance limits are formulated for the Maxwell distribution, and a table of tolerance limit factors for the upper tolerance bound is provided as a function of P (percent of the population below the bound), γ (confidence level), and n (sample size). Values of P, γ, and n considered are P = .50, .75, .90, .95, .99; γ = .75, .90, .95, .99; n = 2(1)25(5)100(10)200(50)300(100)1000,∞.

While the formulation is sufficiently general to be of use to anyone who deals with Maxwell data, examples are restricted to the area of weapon systems analysis. It is in this area that the analyst is often confronted with the problem of estimating the radius of a sphere which will include 100P% of the future burst points from an air burst weapon. Under the assumption that the distribution of burst points about the target center is trivariate normal with common standard deviation σ in all three directions, this development will enable him to attach a confidence statement to the percent of the population encompassed. In particular, it will enable him to determine, on the basis of test firings, the radius of a sphere which will encompass at least 100P% of the future burst points with 100γ% confidence.

3118. Two Bibliographies. For completeness, two bibliographies are mentioned. (02B has copies).

The first is a paper by William C. Guenther and Paul J. Terragno titled "A Review of the Literature on a Class of Coverage Problems," which appeared in the Annals of Mathematical Statistics, 35, 232-260 (1964). This paper considers the multidimensional single-shot kill probability of a weapon against a point target when the impact point of the weapon, the position of the target, and the damage function are given arbitrary probability distributions. The paper lists 58 references, most of which treat special cases of this problem. It reduces these special cases to a standard notation, classifies them, and describes the pertinent aids to computation.

The second source is a bibliography titled "The Coverage Problem," issued as Sandia Corporation Report SCR-443 (Sandia Corporation, Albuquerque, New Mexico, April 1962). The bibliography contains 294 listings, some of which are only indirectly related to the coverage problem proper. These listings are classified under the headings: I. Nomographic Methods of Target

Analysis, II. Related Articles and Publications, and III. Multivariate Normal Integrals and Related Topics. The articles in section I are annotated, those in section II are merely listed, while those in section III are listed and further classified into 12 subsections depending upon the specific subject matter.

PAGE 31-10, REVERSE, BLANK

Section 32

Error Analysis

3201. Introduction. This section discusses these aspects of analysis of error type data.

a. Importance of Axis Components. (3202)

b. Analysis of error data in absolute form in one dimension. Examples are bearing error ignoring sign and miss distance in range ignoring sign. (3203)

c. Analysis of radial error in two dimensions. Examples are radial location error, radial miss distance. (3204)

d. A popular measure of presentation, the CEP (circular error probable). (3205)

3202. Axis/Data Basis

a. In analysis of errors, attention must be paid to the form of the error, the measurement axis, and the meaning of the algebraic sign. If possible, we should strive to have the errors normally distributed in a form that compensates for various test conditions and that has additional operational meaning in interpretation of the algebraic sign. If we are dealing with range error covering a spectrum of ranges from short to long, the error data may be more amenable to analysis (more normal) in ratio vice difference (observed minus true). Bearing error may sometimes be better analyzed in mils or even radians vice degrees. For error in inertia guidance systems, location error per hour since reset may be useful. The axes should be orthogonal and should be operationally meaningful. For example, in aircraft bombing error analysis we do not use the North/South, East/West axis; we use the aircraft flight path and its perpendicular as the axes; the algebraic sign is affixed as to long/short in range and right/left in deflection.

b. If the axis are well chosen and the algebraic signs are available with each error datum, the analysis should be made initially with each axis (or component) separately. For example, in aircraft bombing the errors in range should be analyzed separately from deflection errors. After milking all information thus available, the two components should be combined into other operational measures such as CEP, etc. This is the preferred way for data analysis not only for information on each component that may be useful per se, but also for increased efficiency. Since the analysis based on combining the separate components uses, in effect, twice as much data as, say, treating the radial errors directly, the separate component method is more efficient. For example,

determining the CEP by finding the median of the radials is only 0.46 as efficient as determining the CEP by combining the range and deflection analysis. Our confidence in the CEP determined by the median method is much less (0.46) than the CEP determined by the component method using the same data and sample size. Or in more practical terms, if the CEP derived from combining both components separately is based on say, 30 runs, then for the same confidence we need 65 runs if we use the median of the radials to estimate the CEP.

c. Regardless, in certain projects, instrumentation may be such that the algebraic sign may not be available. The rest of this section discusses data analysis of error data in such cases.

3203. Analysis of Absolute Error, One Dimensional

a. When the error measurements, such as bearing errors, are distributed normally but are recorded without the algebraic sign, the resultant distribution of absolute errors is the folded normal. The positive data include the negative (algebraic) data. This paper, based entirely on reference a, gives tabular values, (extracted) to permit use of standard distribution procedures such as probability of occurrence by magnitude. Table 32-1 relates the observed folded normal parameters to the algebraic normal (unknown) parameters. This permits use of the usual normal distribution tables in most statistical texts. Table 31-2 gives the folded normal distribution directly without recourse to the algebraic normal. Formulae are given in the reference which can be obtained from 02B.

b. The algebraic normal parameters are denoted by μ (mean) and σ (standard deviation). The corresponding folded values have f as a subscript. The fold always occurs at zero on the axis scale. This may or may not be the mean of the algebraic distribution. Different combinations are illustrated graphically in Figure 32-1. The four illustrations give the combinations of small variability (A and C) with large variability (B and D) over the two situations: mean and fold at zero (A and B) with mean and fold not equal (C and D). We can graphically guesstimate the means in each case by visualizing the fulcrums, denoted by Λ. Note that the fulcrums for the high variability cases (B and D) extend further than for the low variability cases (A and C). This points out the lack of independence of of μ, σ when folded.

c. The tables in reference a will be illustrated with an example.

(1) Twenty bearing errors (sign ignored) were obtained in a project. The mean of these absolute data (μ_f) was 14.0 with a standard deviation (σ_f) of 7.79. The μ_f/σ_f ratio, called k, was 1.80.

(2) Entering Table 32-1 with this ratio, we have:

$\sigma_f = 0.9242\ \sigma$

Table 32-1

Values of σ_f and μ by k

k	σ_f	μ
1.4	0.75σ	0.82σ
1.5	0.82σ	1.10σ
1.6	0.87σ	1.30σ
1.7	0.90σ	1.47σ
1.8	0.92σ	1.62σ
1.9	0.94σ	1.76σ
2.0	0.96σ	1.89σ
2.1	0.97σ	2.10σ
2.2	0.97σ	2.13σ
2.3	0.98σ	2.25σ
2.4	0.99σ	2.36σ
2.5	0.99σ	2.47σ

$\mu = 1.6192\ \sigma$

which gives

$\sigma_f = 8.43°$ and

$\mu = 13.6°$

These parameters can be used with standard normal distribution tables (or in this case we can use t-tables because of the small sample size). We can then determine confidence limits, expected probability of meeting some threshold value, etc.

(3) Table 32-2 is to be used with the parameters for the folded normal directly. Areas under the folded normal from 0 to t_f are given for various μ_f/σ_f values. For example, using the 1.8 ratio column entries we note: 10% of the area is less than $t_f = 0.5\sigma_f$ or 3.9° and 25% of the area is above $t_f = 2.5\sigma_f$ or 19.5°. Other typical distribution expectations can be derived.

3204. Analysis of Radials

a. Error analysis in two dimensions is quite common in determining bombing and navigation accuracy. First, as already mentioned in 3202, it is most efficient to analyze each dimension separately and then to combine these results. But if the only data available are in terms of radials, the only recourse is to analyze these radials directly.

FIGURE 32-1

ILLUSTRATION OF FOLDED NORMALS

(FREE-HAND)

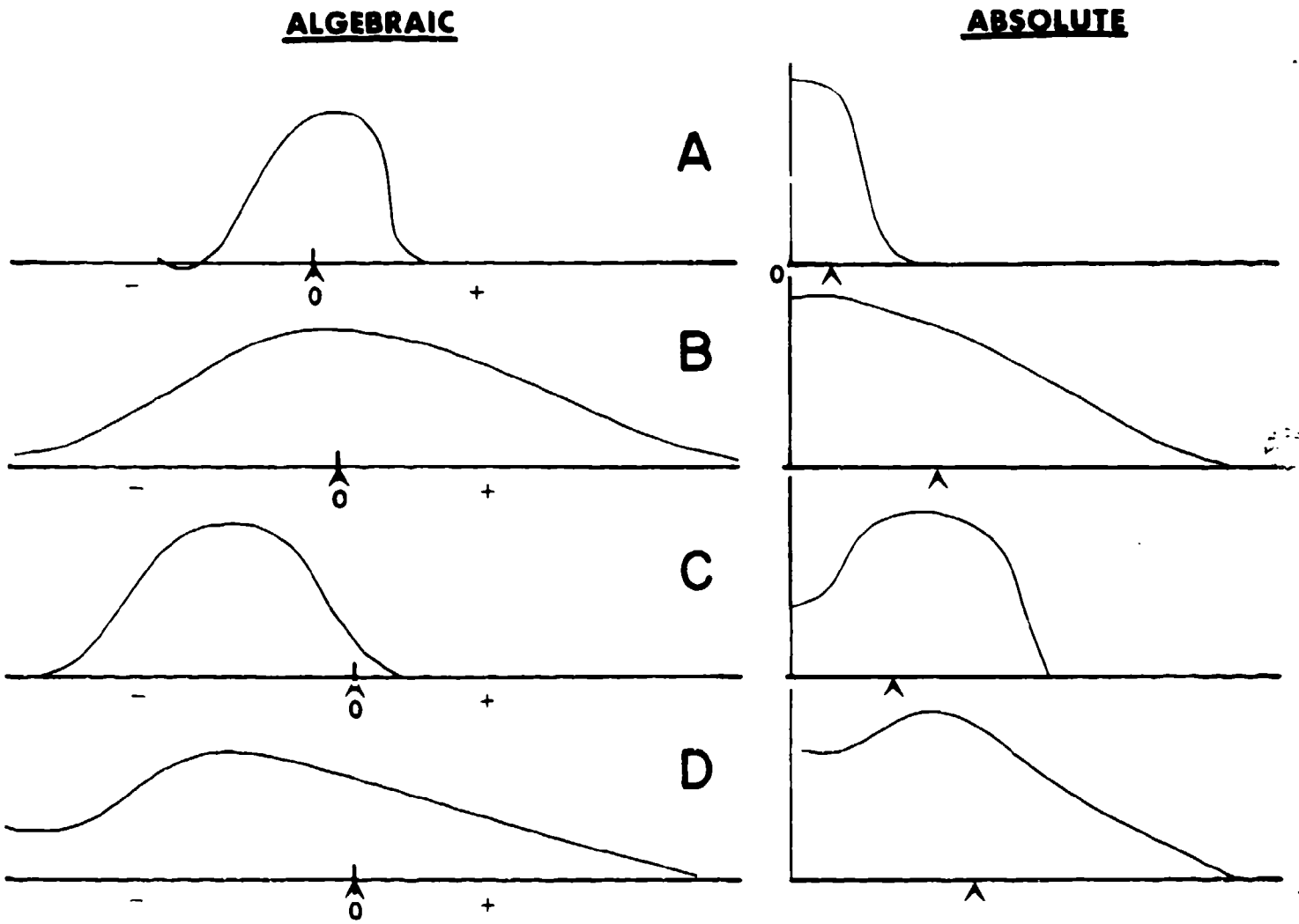


Table 32-2

Area of the Folded Normal by μ_f/σ_f Values

t_f \ μ_f/σ_f	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
0.0	.00	.00	.00	.00	.00	.00	.00	.00	.00
0.2	.09	.06	.04	.03	.02	.01	.01	.00	.00
0.4	.17	.12	.08	.05	.03	.02	.01	.01	.00
0.6	.25	.18	.13	.09	.06	.04	.02	.01	.01
0.8	.34	.25	.18	.13	.09	.06	.04	.02	.01
1.0	.41	.32	.24	.17	.12	.08	.06	.04	.02
1.2	.49	.39	.30	.23	.17	.12	.08	.06	.04
1.4	.56	.46	.37	.29	.22	.16	.12	.08	.06
1.6	.63	.53	.44	.36	.28	.22	.16	.12	.08
1.8	.69	.60	.52	.43	.35	.28	.21	.16	.12
2.0	.74	.67	.59	.51	.43	.35	.28	.21	.16
2.2	.79	.73	.66	.58	.50	.42	.35	.28	.21
2.4	.83	.78	.73	.66	.62	.54	.46	.38	.31
2.6	.87	.83	.78	.72	.66	.58	.50	.42	.35
2.8	.90	.87	.83	.78	.72	.66	.58	.50	.42
3.0	.92	.90	.88	.84	.79	.72	.66	.58	.50
3.2	.94	.93	.91	.88	.84	.79	.73	.66	.58
3.4	.96	.95	.94	.91	.88	.84	.79	.73	.66
3.6	.97	.97	.96	.94	.92	.88	.84	.79	.73
3.8	.98	.98	.97	.96	.94	.92	.88	.84	.79
4.0	.99	.99	.98	.97	.96	.94	.92	88	.84

b. Rayleigh Distribution. Radials, defined as the square root of $x^2 + y^2$, are distributed as a Rayleigh if each dimension, x and y, are orthogonal and distributed normally with mean zero and standard deviation, σ, which is the same for both dimensions. The Rayleigh cumulative is

$$P(r) = 1-\exp(-r^2/2\sigma^2)$$

This gives the probability that the miss distance is less than r. This one-parameter expression is in terms of the individual dimensional distributions. Remember that the means were assumed to be zero and the standard deviations were common for both x and y distributions.

c. The mean ($\bar{r}$) of the radial data is related to the Rayleigh parameter σ, the large sample size relationship is:

$$\bar{r} = 1.25\sigma$$

See reference b, page 15.

The maximum likelihood estimate of σ is:

$$(\Sigma\ R^2/2M)^{1/2}$$

which is used directly with test data to estimate the Rayleigh parameter.

d. Illustration. Suppose we had 30 test runs with position errors in terms of radial distances. The maximum likelihood estimate of σ, using the 30 radials, was calculated to be 480 feet. What is the probability of having a radial error or less than 250 feet?

$$P(< 250)=1-\exp(-250^2/2(480)^2) = 0.13$$

which is the probability of having an error less than 250 feet.

e. Interesting relationships concern particular $K\sigma$ values corresponding to cummulative probabilities:

P	$K\sigma$
.50	1.177
.75	1.665
.80	1.796
.90	2.146
.95	2.448
.99	3.035

f. Reference (c) points out that the use of the above $K\sigma$ is a weak estimate. The confidence with which a circle of radius 1.177 σ contains at least 50% of the population is only .48 even

for a large sample size. To increase the confidence the 1.1774 factor should be increased. For .75 confidence the factor should be 1.28 when the sample size is 20. For .90 confidence the corresponding factor should be 1.38. Below are the tolerance values, Kσ, at 75% and 90% confidence. These are for sample size 20. (The sample size effect is minor.)

P	Kσ	Kσ*	Kσ**
.50	1.177	1.28	1.38
.75	1.665	1.82	1.95
.80	1.796	1.99	2.14
.90	2.146	2.34	2.52
.95	2.448	2.67	2.87
.99	3.035	3.31	3.56

* 75% confidence
** 90% confidence

g. Reference (c) gives an illustration of use of the above Kσ values in the Rayleigh. Suppose ten (N) rounds are fired at a target to obtain the radius of a circle about the target center which will include at least 50% of the future rounds (under similar conditions) from this weapon with 90% confidence. The miss distances from the target in the x (cross range) and y (range) directions are shown below. All measurements are in feet.

x	y
54.7	87.8
-20.1	-178.1
-37.3	-5.6
-136.3	-214.1
-8.1	-23.4
95.8	97.5
91.8	-79.1
116.3	-52.8
-144.5	94.6
75.9	167.3

The σ^2 determination is merely:

$$(\Sigma x^2+\Sigma y^2)/2N$$

Note that the assumption is that σ is common to both x and y and that the means are zero.

$$\sigma = 105.48$$

To estimate weakly (<.50) the radius including 50% of the data, 1.177 σ is used, giving 124. To estimate strongly (.90) the radius including 50% of the data, 1.38 σ is used, giving 146.

h. Note that the Rayleigh is a special case of the Weibull dis-tribution with the Weibull shape parameter equal to 2. This permits use of various Weibull techniques directly in analysis of radials such as Weibull probability graph paper, etc.

i. Circular Error Probable (CEP). The above discussion on tolerance limits with P=.50 is easily recognized as the familiar CEP. Thus, the CEP = 1.177σ or for 75% confidence CEP = 1.28 σ or for 90% confidence CEP = 1.38 σ. (See 3205 for more on CEP.)

j. Maxwell Distribution. (Reference (d)). The Maxwell distribution in three dimensions is similar to the Rayleigh for two dimensions. The assumptions; normality, common σ, zero means are also similar to the Rayleigh.

1. Given the individual (N) trivariate miss distance values from the target center,

$$\sigma^2 \equiv (\Sigma x^2 + \Sigma y^2 + \Sigma z^2)/3N$$

or
in terms of radials

$$\sigma^2 \equiv (\Sigma r^2/3N)$$

2. The tolerance factors have similar functions as for the Rayleigh as follows:

P	Kσ	Ks*	Ks**
.50	1.538	1.65	1.75
.75	2.027	2.17	2.30
.80	2.185	2.34	2.48
.90	2.500	2.68	2.84
.95	2.796	2.99	3.18
.99	3.368	3.61	3.83

* 75% confidence
** 90% confidence

As an example of employing these values, suppose eight rounds(n) are fired at a target to obtain the radius of a sphere about the target center within which at least 50% of the future rounds (under similar conditions) from this weapon will burst with 90% confidence. The radial miss distances from the target center are shown below. All measurements are in feet.

r
201.9
52.9
210.1
120.4

48.1
96.4
85.8
104.7

σ is estimated using j(1) above as $\sqrt{(132130.49)/3(8)}$ = 74.2 ft. To encompass 50% of future round, the spherical radius would be

$$1.538\ \sigma = 114 \text{ ft.}$$

To insure 90% confidence

$$1.75\ \sigma = 130 \text{ ft.}$$

k. Spherical Error Probable (SEP). The above tolerance discussion in three dimensions with P = .50 results in SEP which is similar to CEP's in two dimensions.

3205. CEP

a. A popular summary measure of error analysis in two dimensions, such as aircraft bombing or navigation, is CEP. (circular error probable, circular probable error, circle of equal probability). It is defined as a radius of a circle within which half of the missile impacts or errors are expected to fall. The circle may be centered at the aimpoint or the MPI (mean point of impact). Related to the CEP is the REP and also the DEP. These are the corresponding measures in one dimension. The REP is half the distance between two imaginary points on the ground using an axis along the aircraft approach line equidistant from the aimpoint or expected MPI between which half the missiles are expected to fall. Plus sign usually denotes long in range. The orthogonal axis is related in a similar manner by the DEP, plus meaning to the right.

b. There are various situations and ways to derive CEP. This subparagraph discusses when the standard deviation is assumed to be the same for both dimensions. If this assumption cannot be made, then the analysts should question the wisdom of using a circle to describe an ellipse. When the standard deviation values in each dimension are similar, then as mentioned in paragraph 3204, the large sample size relationship is

$$CEP = 1.1774\sigma$$

or

$$CEP = 0.59\ (s_r + s_d)$$

or

$$CEP = 0.87\ (REP + DEP)$$

This latter is a good approximation even for ellipses when the ratio of the minor to the major axis is greater than 0.2. Note: When the ratio is less than 0.2 and we must find a circle, then use:

$$CEP = REP\left(1 + \frac{DEP^2}{REP}\right).$$

The sample median is a direct estimate of CEP for large samples but as stated in previous paragraphs, it is inefficient. Note that the above assumes MPI is zero after calibration. If this is not the case, we should so state that the CEP is about the MPI or increase the σ estimate by the MPI bias and recalculate the CEP about the aimpoint. The new σ is the square root of the variance about the MPI plus the square of the MPI value. Note that the above CEP estimates should be reduced slightly if based on sample size less than 9.

c. Confidence limits can be estimated on the point estimate of the CEP. These would be the minimum and maximum circles for 50% containment of the expected impacts. If the calculation of the CEP is based on σ, then the Chi Square distribution is used.

$$\text{Interval limit} = CEP\ (\rho/\chi^2)^{\frac{1}{2}}$$

where CEP is the point estimate and χ^2 is based on the confidence level and degrees of freedom, ρ. The degrees of freedom vary by σ estimation procedure. Usually it is the sum of both degrees of freedom going into σ estimation in each X, Y dimension. The degrees of freedom for the elliptical cases are more involved: use

$$1 + K\rho$$

where K is the ratio of the small variance to the larger variance and ρ is the degrees of freedom for the larger variance.

d. For the circular case, reference e graphically presents confidence intervals. See Figure 32-2. Suppose the CEP is estimated based on n=11 data points using the two dimensions. Thus the degrees of freedom p is 2n-2 or 20. The CEP is 90'. Suppose the OTD wishes to know the probability that the true CEP is less than 100'. Enter the graph with the ratio of the two CEPs = 0.9 and p = 20. The probability that the true CEP is less than 100' is 0.70. This figure also shows the impact of sample size. Suppose we are interested in 80% confidence that the observed CEP will be less than 0.8 of the true CEP. Reading the 80% line and the 0.8 line we get p=22 or n=12 in each dimension.

e. If the CEP is obtained using the median of the radials, the confidence interval is obtained indirectly using the binomial. The cumulative binomial with success equal to 0.5 gives the likelihood of obtaining the i-th order observation. For example, if the CEP is derived from the median radial of six impacts, the likelihood of the true CEP being less than the 5th observation (in sequence) is the binomial expansion, summed for 5 out of 6 when the success is 0.5, i.e., the first five terms of the binomial expansion.

f. Circular Error Probabilities. Related to CEP but more important in its own right is the determination of hit for lethal radius $K\sigma_x$ in the following:

Let x and y be two normally and independently distributed orthogonal components of the miss distance, each with mean zero and with standard deviations σ_x and σ_y respectively, with $\sigma_x \geq \sigma_y$.

For various values of $C = \sigma_y/\sigma_x$ Figure 32-3 gives the probability P that the point of impact lies inside a circle with center at the target and radius $K\sigma_x$.

Note: This figure is extracted from reference f. For complete description of the coverage in the reference, see the abstract given in 3112.

g. For completeness with respect to two dimensions having independent normal distributions, the graphical procedure explained in reference e is mentioned. This is the "counting of rectangles" method introduced by Burington, Crow, and Sparugue. This procedure is quite general, say for any target shape, within the normal distribution assumption. (Reference g.)

3206. Reference.

a. Leone, Nelson and Nottingham, The Folded Normal Distribution, Technometrics, Vol 3, No. 4, Nov 1961.

b. Grubbs, F. E., Statistical Measures of Accuracy for Riflemen and Missile Engineers. Webster, RFD#2, Havre De Grace, Md. 21078, 1964.

c. Thomas, M. A., et al Tolerance Limits for the Rayleigh (Radial Normal) Distribution with Emphasis on the CEP. Naval Weapons Laboratory TR-2946, May 1973. Note: See 3116 in Section 31 for scope of tabular values.

d. Thomas, M. A., et al, Tolerance Limits for the Maxwell Distribution with Emphasis on the SEP. Naval Weapons Laboratory TR-2954, June 1973. Note: See 3117 in Section 31 for scope of tabular values.

e. Major Corbisiero, J. V., Evaluating Weapon System Accuracy from a Classical-Bayesian Approach,, Air Force Institute of Technology, June 1979.

f. Hunter, H. T., Circular Error Probabilities, Journal of the American Statistical Association, Vol, 55, Dec 1960, pp 723-731.

g. Crow, E. L., Davis, F. A., and Maxfield, M. W., Statistics Manual with Examples Taken from Ordnance Development, Dover Publications, Inc. 1960, pp 29-30.

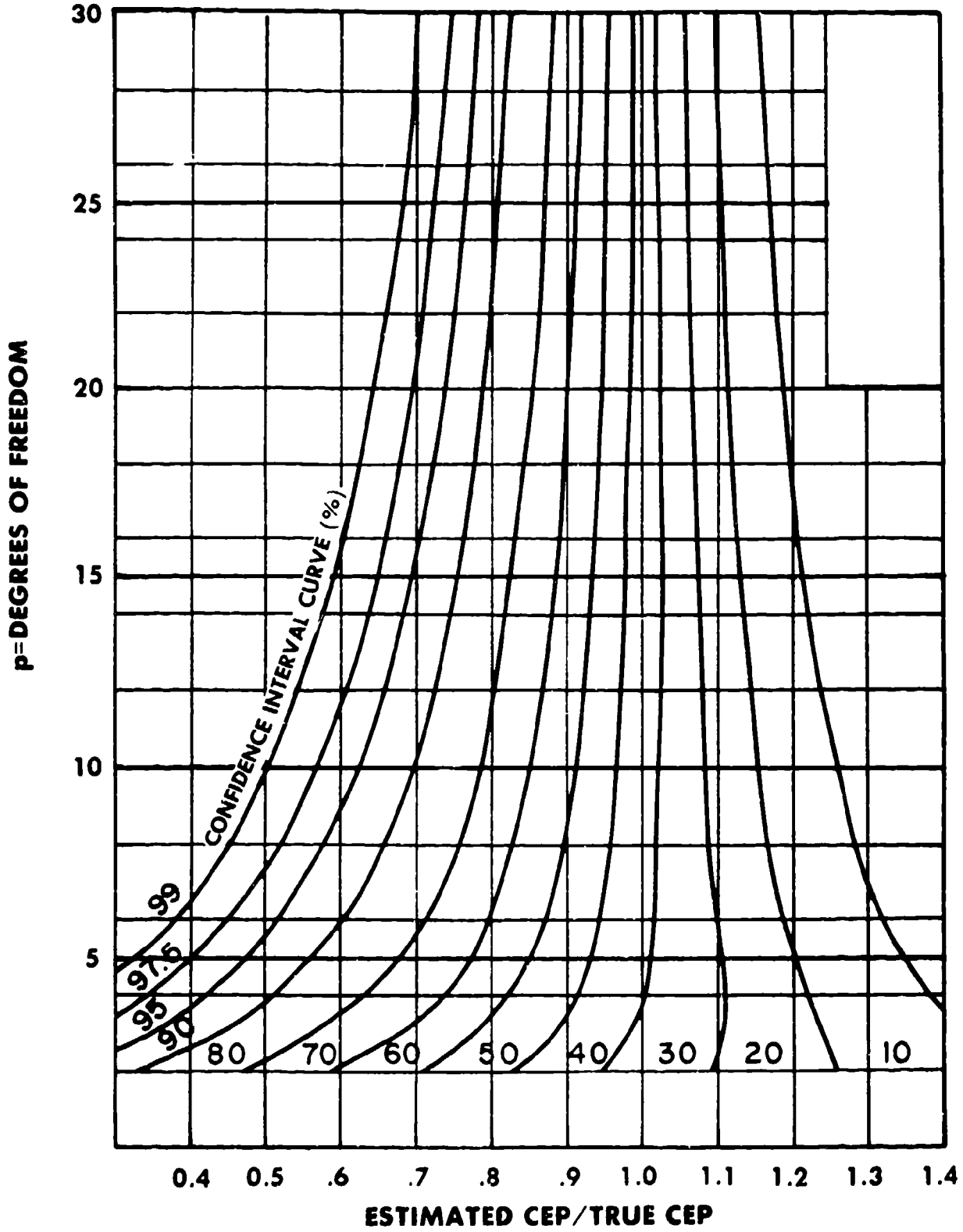
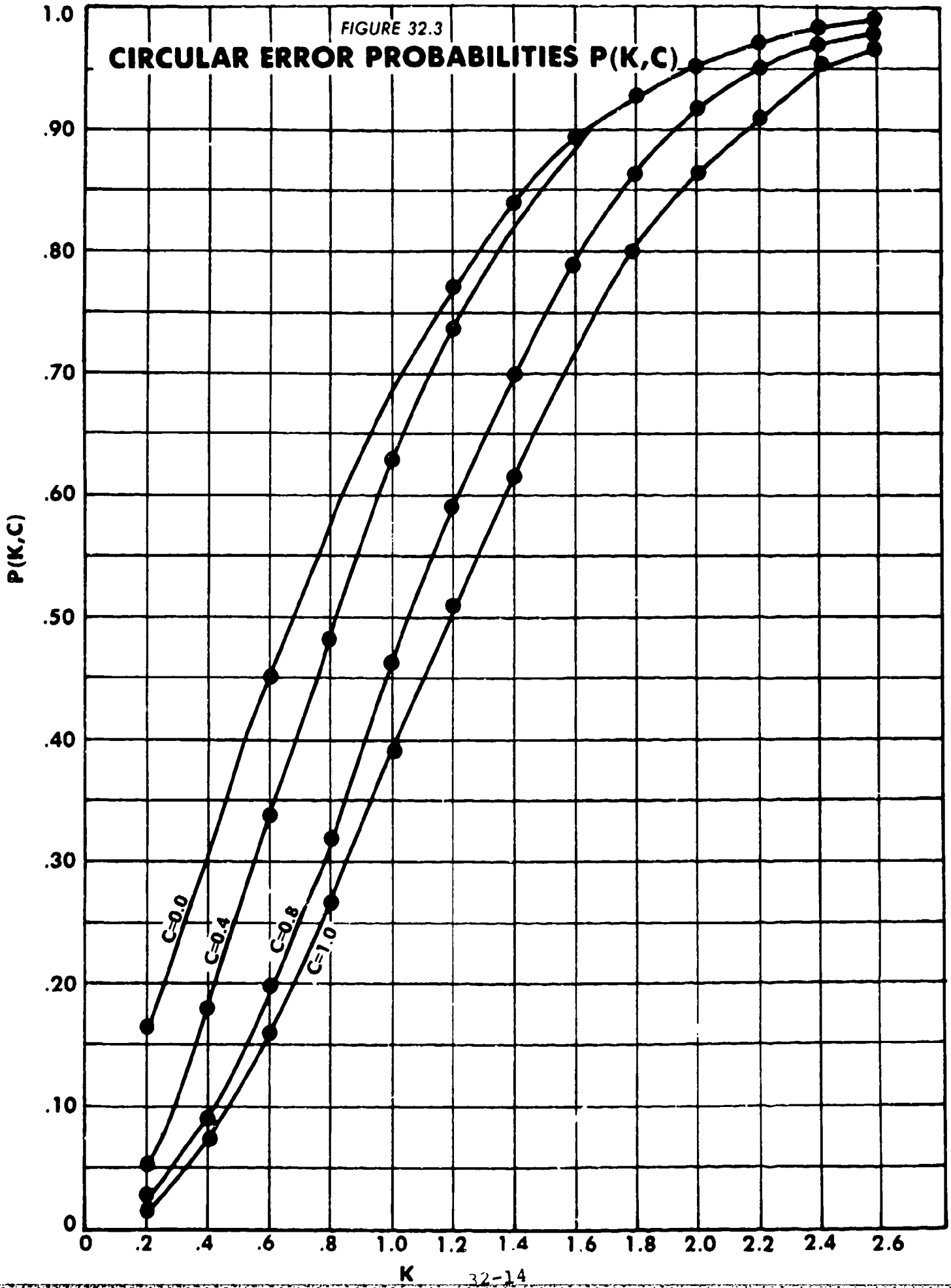


FIGURE 32-2

CHI-SQUARE CONFIDENCE INTERVAL GRAPH

FIGURE 32.3


CIRCULAR ERROR PROBABILITIES P(K,C)
P(K,C)
K
1.0
.90
.80
.70
.60
.50
.40
.30
.20
.10
0
0
.2
.4
.6
.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
C=0.0
C=0.4
C=0.8
C=1.0

Section 33

MTTR or MTTRg?

3301. Introduction. Repair time is an important operational parameter in our evaluations. There are various ways to specify this property. Mean value of repair times is the most prevalent way. This paper focuses on whether this mean should be arithmetic or geometric, i.e., MTTR (mean time to repair) or MTTRg (mean time to repair, geometric).

3302. Distribution of Repair Times

a. Should each set of test data be examined for its apparent statistical distribution? If we have 20 repair times, should we use a statistical procedure to determine whether the underlying distribution is normal, log-normal, or what? The philosophy is that a distribution is a fundamental property of a mechanism, activity, or characteristic vice a particular set of sample data. If a build-up of evidence over the years is available as to a specific distribution, use this as the basis for detection of the inherent distribution. Particularly if it follows some rational explanation. This seems to be the case for repair times.

b. A normal curve, arithmetic scale, extends in theory from negative values through positive values. Since repair times cannot be negative, theoretically, repair times cannot follow a normal curve, arithmetic scale. However, a normal curve, logarithmic scale, is bounded only for positive values. More important, this rationalization is borne out on a practical level with test data. Experience indicates a log-normal. For example MILSTD-471A says:

"The justification for use of the log-normal assumption for corrective maintenance times is based on extensive analysis of field data which have shown that the log-normal distribution provides a good fit to the data."

c. Based on this, the MTTRg is preferred as a parameter of the distribution. Note the careful wording.

d. The actual calculation of MTTRg is directly akin to the normal mean calculation. The observed time values are first put in logarithmic form (let's use the base e). These transformed values are thereafter treated as data. Means, variance, confidence limits, etc. are found with these transformed data using the usual procedures. The very last step is to transform these back to original time units for reporting. (See paragraph 302.)

e. It is interesting to mention the meaning of the antilog of the standard deviations in the log normal case. If the value

is 0.405 in ln terms, the antilog is 1.50. This is not hours. Think relative, i.e., it is 150%. It's 150% of the MTTRg in the following sense. Multiplying and dividing the MTTRg by 150% gives the same results as adding and subtracting 0.405 to the ln MTTRg and then converting to hours.

3303. Feasibility of Using MTTR.

a. While MTTRg is theoretically correct, if the difference between MTTR and MTTRg is not too pronounced, MTTR would be preferred for various reasons. MTTR can be used directly for other purposes than threshold comparisons, i.e., logistic purposes. And it is simple to calculate and explain and use.

(1) MTTRg. This is the mean of repair times on a relative basis. It is usually more representative of repair data. Answers the question: on the average what is the repair time? Closely akin to the median. Useful in decision making, threshold comparisons.

(2) MTTR. This is the mean of repair times on an arithmetic scale. It is strongly influenced by tail or high values. Answers the question of repair rate: if n failures are expected during a cruise, what is my total repair time, clock basis. Useful for workload, manning, logistic determinations.

b. Both measures have a place, answering different questions, and both may be misused. For example, MTTRg should not be used to answer logistic questions.

c. MTTR is always higher than MTTRg. The actual difference between the two is a function of the variation. Theoretically

$$MTTR = \exp (\ln MTTRg + \sigma^2/2)$$

where σ^2 is the true variance of the logarithmic values. Thus, MTTR may be considered "biased" with respect to its use as a threshold value. Note that MTTR is conservative, in favor of the fleet when used as a threshold.

d. Table 33-1 gives the extent of the "bias" or difference between MTTR and MTTRg for various standard deviations. For example, if the standard deviation is 250% (antilog of 0.916) and MTTRg is 2 hours, the expectation for MTTR would be 3.1 hours.

e. Tables 33-2 and 33-3 give the comparison between MTTR and MTTRg for real project data.

f. Some actual project data indicate a wide difference between MTTR and MTTRg. Based on this alone, there is no recourse but to recommend following the theoretically correct measure, MTTRg.

Table 33-1

Theoretical Comparison of MTTRg and MTTR by Size of Variation

Standard Deviation	MTTRg(Hours) 1	2	3	Ratio MTTRg/MTTR
	MTTR Values (Hours)			
150%(1)	1.1	2.2	3.3	0.92
200%(2)	1.3	2.6	3.8	0.78
250%(3)	1.5	3.1	4.6	0.65
300%(4)	1.8	3.6	5.5	0.55

Note: To indicate the variation in data for the various standard deviations, when the MTTRg is 2 hours and the standard deviation is:

(1) 150%, most (90%) of the data will be less than 3.5 hrs.

(2) 200%, most (90%) of the data will be less than 5 hrs.

(3) 250%, most (90%) of the data will be less than 6.7 hrs.

(4) 300%, most (90%) of the data will be less than 9.7 hrs.

Table 33-2

Comparison of MTTR and MTTRg Data
for Some Mechanical Systems

System	n	Normal Dist.		Log Normal Dist.		Ratio
		MTTR (hrs)	Standard Deviation	MTTRg (hrs)	Standard Deviation	MTTRg/MTTR
Oil Water Separator #1	27	1.56	1.68 hrs	0.93	319%	0.60
Oil Water Separator #2	21	1.85	1.53 hrs	1.23	266%	.66
Trash Compactor	8	3.81	2.34 hrs	3.17	195%	.83
Sewage Treatment Plant	13	0.92	0.98 hrs	0.51	374%	.55
Gas Turbine (1)	28	3.59	8.2 hrs	0.91	505%	.25
(2)	26	1.40	1.75 hrs	0.59	401%	.42

(1) Includes two replacements of turbines (32 hrs each).

(2) Excludes the two replacements of turbines, there was another criterion pertain to this property. All of the remaining 26 data were less than 7.5 hrs.

Table 33-3

Comparison of MTTR and MTTRg Data
for Some Combat Systems

System	n	Normal Dist.		Log Normal Dist		Ratio MTTRg/ MTTR
		MTTR (hrs)	Standard Deviation	MTTRg (hrs)	Standard Deviation	
Missile System	24	3.0	3.4 hrs	1.3	605%	0.43
Defensive System	12	2.7	1.6 hrs	2.2	213%	.81
Gun System	14	2.9	5.9 hrs	1.1	403%	.38
Combat System	21	.68	0.59 hrs	0.36	385%	.53

i. Unless and until analysts analyze repair situations to greater depth, MTTR, while desirable, is not feasible.

3304. MTTg or ...?

a. Other measures besides means are useful in requirement specifications; for example, the repair time value to include 90% or 95% of all repair actions at some confidence level (statistical tolerance limits). This type of measure is strongly mission oriented, relating typical mission time to some percentile value. This operational measure is meaningful in a hot-war situation and we should make more use of it.

b. As with all data analysis, extreme care must be taken to avoid averaging "horses and cows." We should deal only with homogeneous groups. Summarize each homogeneous group with means, and then combine these means with weights to get the necessary overall total. (The weight may be external to the testing at sea or depending on the situation.) This search for homogeneous groups and their treatment is an important part of data analysis.

c. For example, MIL-STD-471A mentions systems with built-in diagnostics. Fault-locate times and, therefore, repair times for repairs covered by the built-in test equipment will be different than for other repairs. A plot (or distribution study) of fault-locate times may be difficult to interpret unless such a separation is made.

d. The same argument leads to separation of hardware and software failures. In this situation the correct analysis procedure is reinforced by the need for separate information for parts control, etc.

e. Each project must be carefully analyzed. Separation of major and minor failures? Separation of repairs and replacement of batteries? Separation of repairs and design correction? Etc.

f. In 3302 it was said to use the past as the basis for the proper distribution vice a particular set of operational data. However, this does not mean do not examine the data (greater than 10 in sample size) for distribution. The reason distribution is of interest is not the distribution per se. The primary reason is to determine if the distribution is what it should be. Do repair times follow a log-normal? If not, more analysis may be needed to further separate the "horses and cows."

g. There are many tests for distributions. (Usually a histogram, normal paper plot, etc. is sufficient for our needs.) One of a more recent test for normal (and log-normal) is in Section 35.

3305. $MTTR_g$ Confidence Limits and Tolerance Limits

Suppose confidence limits and/or tolerance limits were needed. The procedures are the usual standard ones for the normal distribution with the added steps of initially transforming the time-to-repair data to logarithms, do not transform back to time-to-repair until the limits are found. See 302 on page 3-2 for an illustrative example.

a. The illustration gives four data:

4.3
6.2
1.8
3.9

which after logarithmic transformation, result in a mean $\bar{x}$ of 1.308 and a standard deviation, s, of 0.520.

b. To get confidence limits say the lower 80%, the normal deviate t based on 3 degrees of freedom, is 0.978. ts is (0.978) (0.520) = 0.509 which, subtracted from the mean 1.308, gives 0.799. The lower 80% confidence limit is 2.22 time units. The $MTTR_g$ is 3.7 time units.

c. The above parameters can also be used to obtain tolerance limits using standard tables (such as Table A-7 in Experimental Statistics, NBS Handbook 91). One such tolerance factor entry for three degrees of freedom is 2.501. The corresponding statement can be made: The probability is 0.75 that at least 90% of the repair times will be greater than $\bar{x}$ -ks = 1.308 - (2.5)(0.520) = .008 or 1.0 time units.

Section 34

Comparison of Two MTBFs

3401. Discussion. A direct way of comparing two MTBFs is to use the property that $2T/\theta$ follow the Chi-square distribution. T is total operating time, θ is true MTBF. (Exponential is assumed.) Mann, et al, point out that if:

$$\theta_0/\theta_1 > F_{1-\alpha}$$

then reject Ho, that two sample MTBFs are the same. F is the usual F distribution. α is the significance level if a one-tail test is warranted. (If not, see 3504.) The degrees of freedom to be used are not firm;, some authors use 2r, some use 2r+1, some use 2r+2, depending on whether the test leads to Ho acceptance or rejection, whether the program is failure-truncated or time-truncated, etc. See for example, Section 10. For conservatism 2r is used here.

3402. Illustration

a. A system was operated for 500 hours with 2 failures during TECHEVAL, and then for 800 hours with 8 failures (during OPEVAL). Was there a significant decrease (one-sided) at the α = 10% level?

b. The F ratio is 250/100 = 2.5 with 4 and 16 degrees of freedom. Standard F tables indicate a significant change ($\alpha \sim .06$).

3403. Reference. Mann, Nancy R, Schafer, Ray E. and Singpurwalla, Nozer D., Methods For Statistical Analysis of Reliability and Life Data. John Wiley & Sons, 1974, p 326.

PAGE 34-2, REVERSE, BLANK

Section 35

Testing for Distribution

3501. Introduction

a. Before any formal tests for distributions are presented, a philosophy point is emphasized by repeating what was said in 3302.a.

"Should each set of test data be examined for its apparent statistical distribution? If we have 20 repair times, should we use a statistical procedure to determine whether the underlying distribution is normal, log-normal, or what? The philosophy is that a distribution is a fundamental property of a mechanism, activity, or characteristic vice a particular set of sample data. If a build-up of evidence over the years is available as to a specific distribution, use this as the basis for detection of the inherent distribution. Particularly if it follows some rational explanation."

Thus, let's admit that testing has gone on for many years, knowledge has been accumulated, and we should use this past accumulation vice deciding on some course of action based on limited (perhaps distorted) data.

b. While we should use the past as the basis for selection of the proper distribution, we should still examine particular sets of data for distribution.

"The reason distribution is of interest is not the distribution per se. The primary reason is to determine if the distribution is what it should be. Do repair times follow a log-normal? If not, more analysis may be needed to further separate the 'horses and cows.'"

3502. Normal Distribution Test

a. Reference a gives a test for normality for small sample sizes (>10). The calculations are straightforward and are presented with an illustration. The test uses a modified version of the Shaprio-Wilk-Francia ω' statistic.

b. Illustration. Eleven (n) data are obtained: 148, 158, 160, 166, 195, 236, 154, 161, 162, 170, and 182.

(1) Step 1: Put data in monotomic sequence. See Table 35-1. Coded y.

(2) Step 2: Get the expected normal order-statistics. Coded m. Many statistical handbooks lists these order statistics by sample size. For example see CRC Handbook of Tables for Probability and Statistics, 1966, page 258.

(3) Step 3: Get the sums indicated in Table 35-1.

(4) Step 4: Determine ω^1

$$\omega^1 = (\Sigma\ my)^2/\Sigma\ m^2\ \Sigma\ (y-\bar{y})^2$$

(5) Step 5: Determine A, B, C values using the formulae in Table 35-1.

(6) Step 6: Determine probability p ≡ exp. C. This is the probability that a ω^1 value this low would be obtained from a normal distribution.

Note: When the sample sizes are large, standard textbooks give procedures.

Table 35-1

Work Sheet for Normality Test

Test Data y	Ordered Normal m	
148	-1.587	$\bar{y} = 172$
154	-1.062	$s = 24.952$
158	-0.729	$\Sigma\ m^2 = 8.884$
160	-0.462	$(\Sigma\ my)^2 = 42617.061$
161	-0.225	$\Sigma\ (y - \bar{y})^2 = 6190$
162	0	
166	0.225	$\omega^1 = (\Sigma\ my)^2/\Sigma m^2 \Sigma (y-\bar{y})^2$
170	0.462	$\omega^1 = 42617.061/(8.884)(6190)$
182	0.729	
195	1.062	$\omega^1 = 0.775$
236	1.587	

$$A = 1.032 - .1836\ (n/10)^{-.5447} = .8577$$

$$B = -.5085 + 2.0768\ (n/10)^{-.4906} = 1.4734$$

$$C = /(\omega^1 - A)/B + .0470)/.0262 - 4.61 = -4.9584$$

Prob = exp C = exp -4.9584) = .007

The probability is very low (.007) that the distribution is normal.

3503. Log-Normal Test. The same procedure for the normal can be used for the log-normal by first converting the individual data to logarithms in Step 1. Then, if "normality" is indicated, the decision really pertains to log-normality.

3504. Exponential Distribution Test

a. There are many methods of testing for the exponential distribution. Gnedenkos' Q test (Gnedenkos, Belyagev, and Salovyev, 1969) has been found to be superior in many aspects. Fercho and Ringer reported this in reference b, which is the basis for this section.

b. The failures are split into two groups, one consisting of the first (earlier) r_1 failures; the other group is the later r_2 failures. r_1 cut-off is arbitary.

$$F = MTBF_1/MTBF_2$$

is distributed as F with $2r_1$, $2r_2$, degrees of freedom under Ho, constant failure rate.

$$MTBF_1/MTBF_2 > F_{\alpha/2,\ (2r_1,\ 2r_2)}$$

then the constant failure rate hypothesis is rejected and it is concluded that $MTBF_1 > MTBF_2$. If

$$MTBF_2/MTBF_1 > F_{\alpha/2\ (2r_1,\ 2r_2)}$$

then the exponential hypothesis is rejected and it is concluded that $MTBF_2 > MTBF_1$. (Note: don't expect much power for total failures numbering less than 15.)

3505. Another Exponential Test

When the number of failures are at least 20, the Bartlett Test is a powerful one for significance testing as to the exponential distribution. The test statistic is:

$$B_r = \frac{2r\left[\ln\left(\frac{\Sigma t_i}{r}\right) - \left(\frac{1}{r}\Sigma \ln t_i\right)\right]}{1+(r+1)/6r}$$

where r is the number of failures and each t_i is the total test time to the i failure. All summations in the above formula are from 1 to r. Note that the remaining operating time after the last failure is not used in the test statistic. The above is straightforward to calculate when one item is being used with repairs. When the evaluation includes more than one item, be

sure to include all test time on all items in determining this test time to the i failure.

b. Under the hypothesis of an exponential distribution. The statistic B_r is chi-square distributed with r-1 degrees of freedom, and a two-tailed chi-square test is in order.

The above is based on reference c.

3506. References

a. Olsson, Donald M. "A Small Sample Test for Non-Normality" Journal of Quality Technology, Vol 11, No. 2, April 1979.

b. Fercho, W.W. and Ringer, L. J. "Small Sample Power of Some Tests of the Constant Failure Rate", Technometrics, Vol 14, No. 3, Aug. 1972.

c. Kapur, K. C., and Lamberson, L. K. Reliability in Engineering Design, Wiley and Sons., 1977 pp 238-246.

Section 36

Screening Seven Variables With Four Runs?

3601. Design of Experiments. Test design, or more formally, the design of experiments, is an important field in the analyst's training. Some designs increase efficiency of testing manyfold. For example the usual way to select test conditions is the piecemeal, one-at-a-time, or baseline approach. Certain designs, if applicable, would furnish the same information with much less firings. An example of this is in Section 24. The key is the difficulty of the operational situation and the variety of situations. Currently, the use of designs in our evaluations is spotty at best;, usually the principles are followed, but seldom are formal test designs used. More work needs to be done in deriving test designs more akin to be operational situation, and on determining the true needs of the operational situation.

3602. Screening. This section presents an extreme test design, called "screening." K variables may be "screened" in N runs when

$$K = 2^{N-1} - 1$$

For example 7 variables can be "screened" in 4 runs! Obviously the assumptions inherent in such a minimum test design must be drastic. These assumptions and an illustration of a "screening" experimental design are presented for impact. The point is that this type is only one class of plans out of many to cover different situations. Application of "screening" test designs to an operational situation directly is too much to expect. However, perhaps some modification may be fruitful. The next three paragraphs present the "screening" test design per se. Possible modifications will be discussed in some section.

3603. Assumptions and Definitions

a. Before discussing assumptions for a "screening" design, remember that there are many unstated assumptions inherent in any developmental program that purports to "test" a new weapon or combat system with, say, 12 firings. Regardless of how the test conditions are selected, a set of assumptions is still inherent in the program.

b. For a "screening" design, K is the number of variables (X_i), each at two levels (-1, +1). These are the number of variables to be tested for importance. However, at most only one of these variables is truly significant. This is the most critical assumption. The analysis determines the variable and the amount of its effect. Δ is the effect of the one true variable on the response data that are normal: mean zero, standard definition σ. The analysis is a simple first-order model.

$$Y = a + bXj$$

The number of runs, N, needed is based on

$$K = 2^{N-1} - 1$$

Monte Carlo results indicate that the size of the error, σ, must be small, about 0.2Δ, relatively speaking.

3604. Illustration: Design. Suppose we have 7 variables X_1, X_2, ...X_7. Each at two test levels. We need

$$7 = 2^{N-1} - 1 \text{ or } 4 \text{ runs.}$$

The specific test conditions for these four runs are formed in two simple steps. The first step is to write to N places the first K binary numbers in ascending order.

Variable	Binary
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

The second step is to replace the zeros with -1. These are the test conditions.

Variable	Binary	Test Conditions RUN			
		1	2	3	4
1	0001	-1	-1	-1	1
2	0010	-1	-1	1	-1
3	0011	-1	-1	1	1
4	0100	-1	1	-1	-1
5	0101	-1	1	-1	1
6	0110	-1	1	1	-1
7	0111	-1	1	1	1

The first column is a run, all variables at low level. The second column is another run with the first three variables at low level, etc.

3605. Illustration: Analysis. Suppose we obtained as the data

Run	Y
1	4.0
2	-2.0
3	4.0
4	-2.0

The analysis is illustrated for the deterministic case (r = 0). (The analysis for the case "with error" is more cumbersome with more steps.) For the error-free case we first find Wi for each run by

$$Wi = \begin{matrix} 0 \text{ if } Yi=Y_1 \\ 1 \text{ if } Yi\neq Y_1 \end{matrix}$$

Run	Y	W
1	4	0
2	-2	1
3	4	0
4	-2	1

Then we find

$$L = \sum_{2}^{N} (Wi)\ (2^{i-2}) = 5$$

This indicates that variable 5 is the one significant variable; this is the one needle in the haystack. (If L = 0, there is no active variable.) The analysis then relates the four test data to the settings for variable 5.

Settings:	-1	1	-1	1
Data:	4	-2	4	-2

We use the first order model

$$Y = a + bX_5$$

The average of all four data is 1 = a.

The coefficient, b, going from -1 to +1 = -12/4 or - 3. So

$$Y = 1 - 3X_5$$

Note: The above is taken from Dr. Smith, who credits OTT and Wehrfritz for this procedure.

3606. Reference. Smith, Dennis E. "Performance Evaluation of a Specific Factor Screening Technique" Desmatics, Inc. Tech Report 113-1. ONR N00014-79-C-0650

Section 37

MTBF Estimate With No Failures

3701. Discussion. In our evaluations, we sometimes observe no failures and cannot directly calculate MTBF. In those cases, we can report a lower one-sided confidence limit for the true MTBF. (See page 10-2.) However, a confidence limit may not be the parameter of interest. In other words there are situations when a MTBF estimate is needed. This paper gives various ways to estimate MTBF when no failures are observed. A recommended way is presented. More important, proper use of this estimate is covered.

3702. Estimates. The fact that the system was stressed for T hours and no failures were observed is important information. While there is no theoretically correct way to determine the "best" estimate of MTBF, we should try to use what we have observed: T hours test time with no failures.

a. Since zero failure rate happens to be the maximum likelihood estimate, some agencies use ∞ as a MTBF estimate. This assumes a failure-free situation.

b. Some agencies use total test time T as the MTBF estimate. This is conservative, as if a failure would occur immediately following T.

c. Another practice is to use the lower 50% confidence limit as the best estimate. This is basically 1.44T, see page 10-2. Reference (a) points out the fundamental difference between confidence limits and the best estimate. But reference (a) also points out the value of good answers, regardless of original purpose of development.

d. Another practice is to use 2T. One basis for this is relating the 50% confidence limit, 1.44T, to the median. The median expectation with the exponential is (MTBF) ln 2. Relating this to 1.44T gives MTBF$\equiv$2T.

e. Reference (a) concludes that 3T is reasonable. This is based in part on extrapolation of the midpoints of the confidence intervals when failures do occur.

f. With the above standard estimation procedures ranging from 1T to 3T (and higher), 2T seems to be a reasonable procedure to use in our evaluations. Thus, 2T $\equiv$ MTBF. However, more critical than the above rule is proper use of the rule, if at all.

3703. Improper Use of 2T $\equiv$ MTBF. The 2T rule is not to be used to determine necessary test time. Nor is it to be used on routine projects except as noted in 3704 below.

3704. Proper Use of 2T ≡ MTBF

a. The rule should be used when a MTBF is needed as part of a more complex procedure. In our use of reliability modeling, say for a complex system, MTBF inputs from operational testing are needed for each system component. When a few of these components under test do not have failures, then the 2T rule can be used so that modeling can procede.

b. For various reasons in an evaluation, the actual test time may be much less than the MTBF threshold. With no failures, two possible types of statements could be made.

(1) Test time was insufficient for any qualitative comparison, i.e., more testing is needed.

(2) Within the limitations of the small test time, satisfactory reliability is indicated.

(3) The 2T ≡ MTBF rule can help in selecting the option (1) or (2) above. If the threshold MTBF is greater than the 2T value, (1) above should be used. If the threshold MTBF is much less than the 2T value, (2) above can be used.

3705. Reference. Welker, E. L. and Lipow, M., "Estimating the expotential failure rate from data with no failure events," 1974 Annual Reliability and Maintainability Symposium.

Section 38

MTBF Testing: 100 Hours With 10 Items Equals 1000 Hours?

3801. Introduction. In testing to determine MTBF, we need a definite amount of test time, say 1000 hours. Is 100 hours with 10 items the same as 1000 hours on one item? After a brief theoretical discussion, this section discusses this from a practical standpoint.

3802. Theoretical

a. The exponential is a paramount assumption in our analysis. This assumes we are past the wear-in stage, but have not reached the wear-out stage. Any failure then depends only on the amount of time under test and the failure rate. The item has no memory. The chance of failure in one hour of testing is the same regardless of when the hour occurred, early or late in the testing period. A failure is due to an internal stress that occurs regardless of the item under test, regardless of being early or late in the testing period.

b. This concept follows human mortality. After the infant matures and before the golden age, there is a long period, say ages 20 to 50, when death is mainly accidental. For this group, the rate can be found from only one year's experience. This concept is related to having only one item under test but repairable. The item is tested, each failure is repaired, and the test continues. There is no hesitation in using combined test times. And rightly so. While we can consider each repair as leading to a "different" item, the item after repair is considered "as good as new." So basically we are in a similar situation as testing many items.

3803. Practical

a. Failures can be due to many things including:

(1) Design.

(2) Quality assurance.

(3) Installation.

(4) Personnel.

(5) Transportation.

(6) Unknown (random).

b. Let's consider these aspects with respect to having one or many items on test. If a failure occurs because of a design flaw, the same flaw is present in each item; one or more items takes no

difference. If a failure could occur because of poor quality assurance, the more items under test the better. If we limit testing to only one item, we may select a lemon or a perfect one; either way, results will be misleading. We need to average out this feature, as we would have in fleet use. The same idea pertains to the other aspects. The more items under test the better. This merely reaffirms the age-old adage; an average is better than a single reading. On the other hand testing many items for short periods may miss important failure modes. So this is a trade-off situation.

c. Dr. Conlon et al in reference (1) talks about this. "As a general rule for evaluation purposes, it is desireable to test a "moderate" number of items for a "moderate" period. A compromise is "... a minimum of three items..." for 1.5 times the threshold or even as high as 3 times the threshold.

d. In summary the trade-off concerns the usual cost in effort, time, money and

1. Likelihood of missing an important failure mode or an important operational stress.

2. Likely item-to-item differences.

Every situation should be examined for the above. Unless 1 above is likely, experience favors many items on test for short periods rather than a few for long periods.

Reference (1): Conlon, J.C. Libius, W.A., Tubbesing, F.H., "Test and Evaluation of System Reliability, Availability, Maintainability - A Primer. Office Director Defense Test and Evaluation, Under Secretary of Defense for Research and Engineering, July 1981.

Section 39

Sample Size For A_o

3901. Introduction. Jan L. Rise in the reference gives α, β formulae for A_o testing. The assumptions do not strictly apply to OT&E, but the results are useable as approximations. The formulae apply to exponentials in both uptime and downtime. While this is not the case in operational evaluations, Rise suggests his formulae are sufficiently robust to be useful. Another limitation is the use of the normal approximation. Rise suggests this is no problem if the evaluation includes at least nine up and down cycles.

3902. Definitions and Formulae

a. A_o is calculated as up over up plus downtime.

b. ϕ is the standardized normal distribution value.

c. A_o^c is the critical A_o value for acceptance.

d. T is the determined test time in mean downtime units that must be estimated so sample size or test time can be determined.

e. A_o' is slightly less than the threshold A_o. A system with true A_o' is a POOR system.

f. A_o° is the producer's quality relevant to the false rejection risk. A system with true A_o° is a GOOD system.

g. α is the false rejection risk. It is the probability of rejecting equipment with a true A_o equal to A_o°. Also called the producer's risk. $1-\alpha$ is the probability of acceptance.

h. β is the false acceptance risk. It is the probability of accepting equipment with a true A_o equal to A_o' or slightly less than the threshold. Also called the consumer's risk.

i. $T = 2\left((\phi_\alpha \cdot A_o^\circ \; 1\text{-}A_o^\circ + \phi_\beta \; A_o' \; 1\text{-}A_o')\; /\; (A_o^\circ - A_o')\right)^2$

j. $A_o^c = A_o^\circ \; A_o' \; (\phi_\alpha \; 1\text{-}A_o^\circ + \phi_\beta \; 1\text{-}A_o') \; /\; (\phi_\alpha \; A_o^\circ \; 1\text{-}A_o^\circ + \phi_\beta \; A_o' \; 1\text{-}A_o'$

3903. Illustration

a. The A_o threshold is 0.80, so $A_o' = 0.8$.

b. We want a false acceptance rate (Navy's risk) of 20%, so $\beta = 0.20$.

c. We want a false rejection rate of 20%. This is $\alpha = 0.20$, the probability of rejection GOOD equipment.

d. GOOD is the producer's quality, which has been set at $A_o^{\circ} = 0.90$.

Note: Only the threshold value A_o' has been furnished directly, say in the TEMP. The other values except T and A_o^c were based on tradeoffs and judgments so typical in this area. T and A_o^c are calculated with the formula already given.

e. With $\phi_\alpha = \phi_\beta = 0.84$ from the normal tables, T is found to be 58.3 mean downtime units. Correspondingly A_o^c calculates to 0.856.

f. The testing procedure to be followed is determined by T and A_o^c. Test for T time; if the observed A_o is higher than A_o^c, accept; otherwise reject.

$A_o^c = 0.856$ and T is 58.5 mean downtime units.

g. Table 39-1 indicates the effect on "sample size" or test time of changing risk factors and quality levels. As expected, the spread between the quality levels has a major part in determining test time.

h. The smallest T value in the Table is 20. This is in terms of mean downtime. Logistic delays play an important part in downtime. Using weights to account for failures not needing off-board supplies, an estimate of downtime <u>per failure</u> is about 200 hours. Using this with the minimum T value, the test time is (20) (200) or 167 days. Very few evaluations have the luxury of almost a half-year of testing. This indicates the nebulous nature of A_o determinations in a typical evaluation. One way around this difficulty is to divide the A_o measure into two aspects: part available on-board and parts (weighted) not available on-board. Section 41 discusses in more detail the aspect with the parts available on board. This aspect doesn't seem to be bothersome. Mean downtime, parts on board, is usually just a few hours; test time is usually just a few hours; test time is usually sufficient for strong estimates for this special case of availability. The logistics delay aspect is difficult to pin down. It varies by fleet and priority so "standard" estimates or sampling procedures should be used with caution.

3904. <u>Reference</u>. Rise, J. L., IEEE 1979 Annual Reliability and Maintainability Symposium Proceedings.

Table 39-1

Variation in Test Time (in mean down time units) by Various Parameters

Threshold A'_o	Producers Quality A_o	False Acceptance β	False Rejection α	Test Time T
.80	.90	.20	.30	40
.80	.90	.20	.20	58
.80	.90	.20	.10	88
.80	.90	.10	.20	97
.80	.95	.20	.20	20
.80	.90	.20	.20	58
.85	.90	.20	20	213
.80	.85	.20	.20	270

Section 40

Confidence Intervals for System Reliability or Effectiveness

4001. Introduction. R (system reliabilty) is sometimes determined from the reliabilities of individual components. Likewise MOMS (measure of mission success) can be determined as a product of individual probabilities. The procedure to get the point estimate is straightforward. However, corresponding confidence limits based on exact solutions are not easily obtained. Excellent approximate methods have been devised. See reference (a). This section concerns using these approximate methods to determine confidence limits. Three cases are presented: confidence limits for a product of binomials, for a series of exponentials, and for a mixture of binomials and exponentials.

4002. Binomial

a. Suppose we wish to determine the lower 80% confidence limit for R of a system of five components in series when each component reliability is a binomial:

$$R = \frac{27}{30} \quad \frac{92}{100} \quad \frac{34}{37} \quad \frac{39}{40} \quad \frac{115}{120} = 0.711$$

b. To get confidence limits, Easterling in reference b proposes to collapse the above into one single pseudo-binomial. Then standard binomial tables or formulae can be used. The pseudo-sample size (N) is first found. This is done by finding the asympotic variance, σ^2:

$$\sigma^2 = R^2 \left(\sum \frac{n_i - s_i}{n_i \quad s_i} \right)$$

where s_i/n_i is the number of successes over the number of trials for the "i" of k components. The summation is over the k components. In the above illustration:

$$\sigma^2 = .711^2 \left[\frac{3}{(27)(30)} + \frac{8}{(92)(100)} + \frac{3}{(34)(37)} + \frac{1}{(39)(40)} + \frac{5}{(115)(120)} \right]$$
$$= .004$$

This is equated to the usual variance formula for the binomial:

$$\sigma^2 = (R)\ (1-R)/N$$

where N is the pseudo-sample size.

$$.004 = (.711)(.289)/N$$

or

$$N = 51.4$$

Using this with the observed R, the pseudo-success number S is obtained

$$S = RN = (.711)(51.4) = 36.6$$

The single pseudo-binomial (rounded upward) becomes 37/52. The usual procedures, such as in reference c, are used to obtain confidence limits, etc. For the illustration, the lower 80% confidence limit is 0.66 using reference c, page 7-3.

c. Winterbottom in reference d, modifies the above procedure. Rather than calculate N, Winterbottom takes the smallest denominator as N. The rest of the procedure is unchanged. In the five component illustration, N would then be 30. Using this, S would be (.711)(30) = 21.3, which rounds to 21. So the pseudo-binomial is 21/30. Using reference c, the lower 80% confidence limit for R is 0.61.

d. The Winterbottom procedure is conservative; i.e., there is a tendency to have lower 80% confidence limit determinations (since N is minimum). Easterling is the preferred procedure except for indeterminate cases, such as no failures in each component, i.e., R = 1.0.

4003. Exponential

a. To obtain a system MTBF when the data are available only for components in series, the sum of the individual component failure rates are used. (Exponential is assumed.) This is a simple procedure. To get the 80% confidence limit, however, approximate methods must be used. For example, the system lower 80% confidence limit could be found by combining the lower P% confidence limits of the components. This paper heuristically derives what the P% confidence limits of the components should be to end up with 80% system confidence limits. (The extension to other than 80% is straightforward.) The premise is that the system lower confidence limit should be the same when based on combining the component limits or when based on the system directly.

b. When based on the system directly, the standard procedure is to use:

$$2T/\chi^2_{2r+2} \quad \ldots\ldots (1)$$

where T is the total test time and r is the number of failures. If T was 1000 hours with two failures, the 80% lower confidence limit would be 234 hours. If the situation was viewed as two (k) components in series, the test time would still be 1000 hours and in the ideal case the failures would be equally distributed, $r_1 = r_2 = 1$ each. For the system value to be 234 hours, each of the two components must be 467 hours. That is, combining failure rates, two 467 hour values combine in series to 234 hours. The corresponding χ^2_{2ri+2} value to get 467 hours is 4.28 which is 1/2 the χ^2_{2r+2}

value (8.56) used to obtain the system limit of 234 hours using equation 1.

c. When the number of failures are equally distributed among the components, the χ^2 term in equation (1) is found by:

$$\chi^2_{2ri+2} = \frac{1}{k} \chi^2_{2r+2} \quad \ldots\ldots (2)$$

where the left-hand term (to be derived) pertains to individual components and the right-hand term pertains to the system. k is the number of components.

d. Table 40-1 is a two-part table for system χ^2 and is presented for the record. The columns are for k = 2, 3, 4, and 5. The rows are for r_i = 1, 2, 3, 4, 5, and 10. These rows are per component. So for k = 3, r_i = 2 the system number of failures, r is 6. Table 40-1a is for 2r + 2 degrees of freedom. Table 40-1b is for χ^2_{2r+2} for lower 80% confidence for the corresponding degrees of freedom.

e. Table 40-2 pertains to individual component values. For the same k and r_i values, Table 40-2a gives the solution, left-hand side of equation (2). This is the entry to use with equation (1) to derive the 80% confidence limit for the system.

f. Illustration: Suppose we had a three component system. The test time was 1000 hours each. Five failures were observed with each component. k = 3, r_i = 5 entry in Table 40-2a is 12.83. Using equation (1), the value per component is 155.9 hours. The failure rate is .0064. With three components, the system failure rate is .0192. This converts to 52 hours, the system lower 80% confidence limit for MTBF.

g. Suppose the 15 failures were distributed unequally, such as 3, 2, 10 for the three components. Table 40-2a entries are still used with a similar procedure.

Component	A		B		C		System
T	1000		1000		1000		
r_i	3		2		10		
Table 40-2a	8.33		6.07		23.70		
Value (hours)	240.1		329.5		84.4		
Failure Rate	.0042	+	.0030	+	.00118	=	.0190

Conversion of system failures rate gives 52 hours as before.

h. Suppose test times were not equal for components. Suppose in the preceding, T for C component was 500 hours rather than 1000. Still using the Table 40-2a entry, the equation (1) value is 42.2 hours for this component. Combining the three components now results in a system confidence limit of 32.4 hours, which appears reasonable.

i. Table 40-2b is presented as a matter of interest. Entries in Table 40-2a were used with standard χ^2 tables. Table 40-2b gives the percentage per component to result in a system 80% lower confidence limit. Thus, the 61% value should be used when k = 4 and r_i = 10.

4004. Mixture: Binomial and Exponential. In the MOMS calculation, the individual entries are mainly binomial type. One or two components in such a series may have the respective reliabilities observed via the exponential. We can translate these cases to a binomial. Example: suppose the reliability of a component is based on a 1500-hour test with three failures. The observed MTBF is 500 hours, the corresponding lower 80% confidence limit for MTBF is 270 hours. If the mission time is 50 hours, the MTBF values give the following reliabilities:

$$\exp(-50/500) = 0.90$$
$$\exp(-50/270) = 0.831$$

Reference e is scanned looking for 0.83 as the lower 80% confidence limit and three failures. Usually one or two binomial measures meet the match. Then the selection hinges on matching the point estimate or reliability. In this case looking at the 80% level column in reference e and matching the 0.83 value with 3 failures give 28/31. This is checked since 28/31=0.90.

4005. References

a. Mann, Nancy R. and Grubbs, Frank E. Approximately Optimum Confidence Bounds for System Reliability Based on Component Test Data, Technometrics, Vol. 16, Number 3, August 1974. (with 29 references)

b. Easterling, Robert G., Approximate Confidence Limits for System Reliability, In American Statistical Association, March 1972.

c. Natrella, Mary Gibbons, Experimental Statistics, NBS Handbook 91, August 1963.

d. Winterbottom, Alan, Confidence Limits for Series System Reliability from Binomial Subsystem Data, In American Statistical Association, September 1974.

e. Cooke, James R.; Lee, Mark T.; Vanderbeck, John P., Binomial Reliability Table (Lower Confidence Limits for the Binomial Distribution) NOTSTP 3140, January 1964. From DTIC AD444344. This gives lower one-sided confidence limits for 80%, 90, 95, 97.5, 99, and 99.5 for N = 4(1)500.

Table 40-1

SYSTEM Work Tables

a: χ^2 Degrees of Freedom (2r+2)

Failures Per Component (ri)	Number of Components (k)			
	2	3	4	5
1	6	8	10	12
2	10	14	18	22
3	14	20	26	32
4	18	26	34	42
5	22	32	42	52
10	42	62	82	102

b: Corresponding χ^2 Values for 80% Lower Confidence Limit

ri	k			
	2	3	4	5
1	8.6	11.0	13.4	15.8
2	13.4	18.2	22.8	27.3
3	18.2	25.0	31.8	38.5
4	22.8	31.8	40.7	49.5
5	27.3	38.5	49.5	60.3
10	49.5	71.1	92.5	113.9

Table 40-2

Component Work Tables

a: χ^2 Values for 80% Lower Confidence Limit

ri	k			
	2	3	4	5
1	4.28	3.67	3.35	3.16
2	6.70	6.07	5.70	5.46
3	9.10	8.33	7.95	7.70
4	11.40	10.60	10.18	9.90
5	13.65	12.83	12.38	12.06
10	24.75	23.70	23.13	22.78

b: Percentages Values for 80% Lower Confidence Limit

ri	k			
	2	3	4	5
1	63	55	50	47
2	65	58	54	51
3	66	60	56	54
4	67	61	57	55
5	68	62	58	56
10	69	64	61	59

Note: Linear interpolation used throughout.

Section 41

Confidence Limits for A_o (No Logistics Delay)

4101. A special case of A_o (operational availability) is when there can be no logistics delay causing downtime. This is the case for software availability. It is also the case when cannibalization is used or when local supplies carry a complete inventory. This special case availability is

$$A_1 = \frac{MTBF}{MTBF + MRT} \quad \ldots\ldots\ldots\ldots (1)$$

when MTBF (mean time between failure) pertain to mission-aborting failures.

MRT (mean restore time) is active repair time plus time to get parts from local sources. Gray and Lewis in the reference gives a way to determine confidence limits for A_1. This section gives the conditions and tabular values to determine the lower 80% confidence limit.

4102. Conditions:

$$\frac{\theta}{\theta + \mu} = \frac{1}{1 + \frac{\mu}{\theta}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

when θ is distributed exponentially and μ is distributed log normal. For many situations this ratio can be translated directly to A_1.

a. When the system under evaluation is to be used continuously in the Fleet, then the testing at-sea, to be representative, would be continuously energized except for downtime. Uptime then would be a direct function of MTBF or θ.

b. Likewise μ can be taken as the average downtime. This includes repair time, MTTR and parts procurement time, local sources.

c. An important assumption is that the variance v of the log normal downtimes is known (or is determined based on a large sample). Section 33 gives real data estimates for the repair part of the downtime. See pages 33-4 and 33-5. The median is 375% standard deviation or 1.32 in log e terms. The variance is 1.75 (or 575%) for active repair time. Parts procurement time, local, will broaden the MTTR to MRT (mean restore time). This more operational measure, usirg MRT, will have slightly more variance than for MTTR alone. The variance estimate is arbitrarily taken as 2.0 in ln terms (or 740%). This value will be used later in paragraph 4104.

4103. General Formula

The formula for confidence limits about the ratio is:

$$\frac{2\ m\ MTBF}{2\ m\ MTBF + b\mu\ \exp(v/2)} \quad \ldots\ldots(3)$$

where m is number of failures
μ is mean downtime (geometric mean) in same time units as MTBF.
v is variance of downtime in ln to base e
n is number of downtime events; needed to find b
b is parameter tabulized in the reference.

The reference gives the b parameters corresponding to 0.05, 0.10, 0.25, 0.50, 0.75, 0.90 and 0.95 confidence. Linear interpolations were made for 0.80 and are given in Table 41.1. These values are to be used to obtain the lower 80% confidence limit.

Illustration: Suppose 30 downtime events resulted in a geometric mean restore time of 2.355 hours and a variance of 1.603 (still in ln). Suppose also that the reliability test resulted in a MTBF of 212 hours based on 10 failures. The A_1 using (1) above calculates to be 0.9890. To determine the lower 80% confidence limit for A_1, we need b from Table 41.1. To read this table n/v = 30/1.6 = 18.75. Using this in the m = 10 column we interpolate and get b = 27.1. Using this in formula (3) we get 0.967 as the lower 80% confidence limit about the observed A_1 value of 0.989. Note that the variance of downtimes was assumed to be the calculated value, based on 30 events in this case.

4104. Specific Estimates

The usual situation in project evaluations is for small sample sizes. Suppose there were ten failures and both the uptime and downtime values were to be based on these ten failures. Let's say the MTBF was 212 hours and the mean downtime was 2.355 hours. Rather than calculate the variance v, the estimate arbitrarily taken as 2.0 in paragraph 4102 should be used. The formula (3) simplifies to

$$\frac{2\ m\ MTBF}{2\ m\ MTBF + 2.72\ b\mu} \quad \ldots\ldots(4)$$

b in Table 41.1 is 31.2. The 80% lower confidence limit using (4) is 0.955 about the observed A_1 value of 0.989.

4105. Reference

Gray, H. L. and Lewis, T. O.," A Confidence Interval for the Availability Ratio," Technometrics, Vol. 9, No. 3, August 1967.

Table 41.1

Values of b for Lower 80% Confidence

n/v	m(number of failures)								
	5	6	7	8	9	10	11	12	13
5.0	16.2	19.2	22.2	25.2	28.2	31.2	34.2	37.2	40.2
6.0	15.8	18.7	21.5	24.5	27.4	30.3	33.2	36.1	39.0
6.7	15.6	18.4	21.3	24.2	27.0	29.9	32.7	35.6	38.4
7.5	15.4	18.2	21.0	23.8	26.6	29.4	32.2	35.0	37.8
8.0	15.3	18.1	20.9	23.6	26.4	29.2	32.0	34.7	37.5
10.0	15.0	17.7	20.4	23.1	25.8	28.5	31.2	33.8	36.5
15.0	14.5	17.2	19.8	22.4	24.9	27.5	30.1	32.8	35.2
20.0	14.3	16.9	19.4	22.3	24.5	27.0	29.5	32.0	34.5
30.0	14.1	16.6	19.1	21.6	24.0	26.5	28.9	31.3	33.7
40.0	14.0	16.5	18.9	21.4	23.8	26.2	28.6	31.0	33.3

Index